2017

Agent Intermediation and Racial Price Differences

Adam Nowak
West Virginia University, Adam.Nowak@mail.wvu.edu

Patrick S. Smith
San Diego State University, patrick.smith@sdsu.edu

Department of Economics

Working Paper Series

Agent Intermediation and Racial Price Differences

Adam Nowak

Patrick Smith

Working Paper No. 17-21

This paper can be found at the College of Business and Economics Working Paper Series homepage:

http://business.wvu.edu/graduate-degrees/phd-economics/working-papers

Agent Intermediation and Racial Price Differentials

Preliminary Draft - Please Do Not Cite

Adam Nowak *
West Virginia University

Patrick Smith †
San Diego State University

September 7, 2018

Abstract

Most housing transactions are brokered wherein the buyer and seller do not meet in person. In which case the buyer's race is not revealed to the seller, so the seller cannot discriminate based on race. Despite this observation, previous studies find racial price differentials based on the race of the buyer. We provide evidence that these estimates suffer from an omitted variable bias attributable to the time-varying attributes of the house. After controlling for the time-varying attributes of the house, we find that minority (black and Hispanic) and non-minority (white) buyers pay a similar price for comparable housing. We also examine whether agent intermediation provides a channel through which differential treatment can occur. We find no evidence of racial price differentials at the agent level.

Key Words: Housing, agent intermediation, discrimination, racial price differentials

*West Virginia University; Email: adam.d.nowak@gmail.com
†San Diego State University; Email: patrick.smith@sdsu.edu

[1]We would like to thank Valentino Demarco, Nadia Greenhalgh-Stanley, Crocker Liu, Blair Russel, Will Strange and seminar participants at Arizona State University, the 2017 Urban Economics Association, 2018 American Real Estate Society, and 2018 AREUEA National meetings for helpful comments. All errors are the responsibility of the authors.

1 Introduction

We find that minority (black and Hispanic) and non-minority (white) buyers pay a similar price for *comparable* housing. This finding conflicts with a rich extant literature that finds evidence of racial price differentials in housing markets.[1] First, we replicate the racial price differentials reported in the extant literature using data from Atlanta, Georgia. Next, we show that the repeat-sales approach employed in Myers (2004) and Bayer et al. (2017) suffers from an omitted variable bias stemming from the time-varying attributes of the house. Although house fixed effects control for the time-invariant attributes of the house, they do not control for the time-varying attributes. Thus, the repeat-sales pairs are not necessarily *comparable* housing.

We control for the time-varying nature of the housing stock using two distinct approaches. First, we use a series of filters to remove repeat-sales transactions that were purchased by investors and/or had a short holding period. Filtering out these transactions addresses concerns that the houses were purchased and rehabbed (i.e. flipped). Second, we augment the repeat-sales approach with textual information about the time-varying attributes of the house. After we employ the filters/textual analysis, the racial price differentials in the extant literature become statistically insignificant indicating that renovations and improvements to the condition of the house are confounding factors.

The textual analysis we employ uses a data-driven methodology that does not rely on a predefined word list or dictionary.[2] Instead, we use a variable selection procedure for high-dimensional data that selects relevant words and phrases (tokens) present in the listing agent's description of the house. We then include these tokens directly in the repeat-sales

[1] King and Mieszkowski (1973), Yinger (1978) and Bayer et al. (2017) find that minorities pay a premium for housing. In contrast, Chambers (1992) and Kiel and Zabel (1996) find that minorities pay less for housing. We discuss these conflicting results and provide a detailed review of the literature in Section 3.

[2] The application of textual analysis in academic research has increased significantly over the past decade. The overwhelming majority of the studies that employ textual analysis in the economics and finance literature use a predefined word list or dictionary craeted by the researcher. Recently, Baker et al. (2016) use a predefined word list to create an index of economic policy uncertainty, Loughran and Mcdonald (2011) construct a finance-specific dictionary.

estimation. The procedure advances the methodology in Liu et al. (2018) to identify and delineate between time-invariant and time-varying tokens in unstructured text. Using the enhanced procedure, we provide evidence that repeat-sales estimates of racial price differentials are subject to an omitted variable bias stemming from the time-varying attributes of the house. More specifically, we show that including the time-varying tokens in a repeat-sales model mitigates the bias associated with transactions in which minority buyers purchase houses that were recently rehabilitated, renovated, or otherwise improved by non-minority sellers.

To the best of our knowledge, this is the first study to document the absence of racial price differentials in housing markets. Several controls, which we discuss in the ensuing subsections, are in place to prevent racial price differentials, so the lack of price-based discrimination is to be expected.[3] This finding does not, however, imply there is no discrimination in housing markets as minorities may pay a similar price in certain locations, yet face difficulties when purchasing houses in other locations (i.e. steering).

1.1 Agent Intermediation

In most housing transactions the buyer and seller do not meet or communicate prior to the closing - if at all. Instead, each party is represented by a real estate agent who acts on their behalf. This agent intermediation should, in theory, eliminate racial price differentials in housing markets. If the seller never meets the buyer - discrimination is not feasible. The lack of racial price differentials supports this conjecture, but does not rule out other forms of racial discrimination - particularly at the agent level.

In-person fair housing audits and online correspondence studies show that market intermediaries are a source of differential treatment in housing markets (Yinger, 1986; Ewens et al., 2014). As such, we examine whether the pervasive use of real estate agents provides a

[3]Similar to Yinger (1995) we define discrimination as adverse treatment of an individual based solely on her or his membership in a particular racial or ethnic group.

mechanism through which differential treatment occurs.[4] In doing so, we shed light on the following questions: If the buyer and seller never meet, how is the buyer's race revealed? Does the race of the buyer's agent serve as a proxy for the race of the buyer? If so, can a minority buyer avoid discrimination by hiring a non-minority buyer's agent? These questions have several important policy implications that are directly associated with the U.S. government's fair housing policy.

Fair housing has long been a central focus of governmental policy. The Fair Housing Act (FaHA) of 1968 prohibits discrimination in housing on the basis of race, color, national origin, religion, sex, disability, and the presence of children. In 1988, the Fair Housing Amendments Act (FaHAA) was enacted to help enforce FaHA. Among other things, FaHAA removed the limit on punitive damages, lengthened the statute of limitations, and created a system of administrative judges to handle claims of discrimination. Although considerable progress was made in the late 20th century, there is still evidence of racial discrimination in housing markets (Ahmed and Hammarstedt, 2008; Hanson and Hawley, 2011; Ewens et al., 2014). We argue that identifying and understanding the channels through which discrimination may occur in housing markets can strengthen and aid in the enforcement of the fair housing policies that are already in place.

1.2 Agent Representation

Although agent intermediation should, in theory, eliminate racial price differentials, it could also be a the channel through which price-based discrimination occurs. Despite this conjecture, real estate agents role in price-based discrimination has, for the most part, been ignored in the extant literature. For example, Bayer et al. (2017) note that "the arms-length nature of agent/broker-facilitated transactions may limit the information the homeowner has about a potential buyer's race and ethnicity", but do not examine the agent/broker channel in their

[4]Real estate agents were involved in approximately 88% of all housing transactions in 2016 (NAR, 2016a). Throughout the paper, we use the term real estate agent to refer to all salespeople and brokers. We use the term listing agent to refer to the agent representing the seller and the term buyer's agent to refer to the agent representing the buyer. We discuss alternative agency relationships in an internet appendix.

analysis. We examine this channel by collecting and classifying the demographic information for 4,906 real estate agents in Atlanta, Georgia. To the best of our knowledge this is the first study to examine agent representation by race and estimate racial price differentials at both the buyer and agent level.

We provide evidence that buyers sort not only into neighborhoods based on race, but also in terms of agent representation. Minority (non-minority) buyers are disproportionately represented by minority (non-minority) agents relative to the underlying real estate agent population. Buyers' propensity to select agents of a similar race provides a mechanism for racial discrimination at the agent level through two distinct channels. First, the race of the buyer's agent may serve as a proxy for the race of the buyer in negotiations with the listing agent and seller. Second, the buyer's agent is the only party in the transaction that almost certainly knows the buyer's race, so the buyer's agent may be the mechanism through which differential treatment occurs.

Given the racial sorting we document, we test whether price differentials exist at the agent level. If the race of the buyer's agent serves as a proxy for the race of the buyer and white listing agents (sellers) discriminate, it is possible that buyers represented by black agents pay a premium relative to buyers represented by white agents. This, however, is not the case. We find that buyers represented by black and white agents pay a similar price for comparable housing. Although we find no evidence of racial price differentials at the buyer or agent level, we do not interpret these findings as an absence of any racial discrimination in housing markets. One concern is that non-minority sellers (listing agents) may prefer not to work with or are less willing to negotiate with minority buyers (buyer's agents).[5] In which case, the racial price differentials we report do not capture overt taste-based discrimination that may be present in housing markets.

Statistical discrimination is another form of racial discrimination that may not be cap-

[5]A recent survey of individuals that bought a house or tried to buy a house in 2016 finds that nearly half (48.8%) of the 569 minority buyers felt that sellers or their agents were less eager to work with them because of their race (Scharnhorst, 2017).

tured when testing for racial price differentials in housing markets. Ondrich et al. (2003) find evidence of statistical discrimination that they attribute to an agent's uncertainty about black buyers' ability to put forth successful bids. We examine whether this form of statistical discrimination has an effect on the price minority buyers pay for housing using data that identifies preapproved buyers. Preapproved buyers receive a written commitment from their lender stating that they will extend a home purchase loan up to a specified amount. The preapproval letter should negate the agent uncertainty documented in Ondrich et al. (2003). However, we find that controlling for whether the buyer was preapproved has no discernible effect on the price minorities pay for housing.

1.3 Indirect Financing Controls

Although the bulk of our discussion and analysis focuses on agent intermediation, we recognize that racial price differentials may also be limited by indirect controls tied to mortgage financing. To obtain financing a third party appraisal is required for most residential loans. If the appraised value is less than the agreed upon sales price, the buyer has to either (i) increase their down payment to cover the difference, (ii) renegotiate with the seller, or (iii) cancel the deal. Regardless of which option the buyer chooses, the appraisal provides the buyer with an estimate of the house's market value and offers an escape clause. Real estate agents are aware of this fact, so it should limit the size of the racial price differentials observed in the market. Of course, this raises concerns that there are unobserved instances in which the racial price differential was so high that the minority buyer could not afford or decided not to purchase the house. In which case, racial price differentials associated with overt taste-based discrimination are censored from the estimates we report.

The remainder of the paper is organized as follows. Section 2 describes the dataset employed in this study. Sections 3 and 4 review previous studies of racial discrimination in housing markets, highlight their limitations, and discuss the methodology we use to overcome the limitations. Section 5 presents our findings and Section 6 offers our concluding remarks.

2 Data

This study focuses on the single-family detached housing market in Atlanta, Georgia from January 2000 through September 2016. The study area includes the five counties (Clayton, Cobb, DeKalb, Fulton, and Gwinnett) that form the core of metro-Atlanta. Metro-Atlanta is an ideal location for this study given its size and racial diversity. According to the 2010 Census, the Atlanta metropolitan statistical area (MSA) was the ninth largest MSA in the United States and the City of Atlanta was the fifth largest black-majority city.[6]

The dataset employed in this study draws from several sources: CoreLogic, Home Mortgage Disclosure Act (HMDA), and Georgia Multiple Listing Service (GAMLS). The CoreLogic data includes every real estate transaction recorded in the five counties' tax assessor offices from January 2000 through September 2016. The dataset includes the transaction date, transaction price, deed type, lender name, and loan amount for each transaction. It also includes a unique identifier for each house, sale date, sale price, and many (nearly) time-invariant house level characteristics such as square feet living area, lot size, number of bedrooms, and number of bathrooms.

We match demographic information about the buyer to each transaction in the CoreLogic data using the publicly available loan application registry (LAR) data gathered under HMDA. The demographic information includes the buyer's race and ethnicity. We merge the LAR and CoreLogic data using the following fields: census tract, transaction year, lender name, and loan amount. We discuss the merge process in more detail in the appendix.

The GAMLS data includes house characteristics available in the CoreLogic data as well as transaction details (sales price, time-on-market, etc.), and an unstructured written description (remarks) about the house for every listing from January 2000 through September 2016. We describe how we transform the unstructured text in the remarks into regressors in Section 4.3. The GAMLS data also identifies the listing agent and buyer's agent involved

[6]Detroit (MI), Memphis (TN), Baltimore (MD), and Washington (DC) were the only black-majority cities with a larger black population.

in the transaction. We merge the GAMLS and CoreLogic-LAR data using the following fields: property address, sales date, and sales price. Additional information about the merge process is provided in the appendix. Prior to merging the three datasets we apply several filters to clean the data. A detailed list of the filters and descriptive statistics for the full raw dataset are provided in the appendix.

For the repeat-sales estimates, we drop transactions for any house that sold only once during the study period. Table 1 displays descriptive statistics for the CoreLogic and GAMLS repeat-sales samples in Panels A and B, respectively. Column 1 in both panels includes every repeat-sales transaction in which we know the race of the buyer. Columns 2 to 4 are filtered by neighborhood racial composition. Column 2 includes repeat-sales transactions in neighborhoods (i.e. census block groups) that were less than 50 percent white according to the 2010 Census. Columns 3 and 4 include repeat-sales in neighborhoods that are greater than 50 percent and 80 percent white, respectively. The neighborhood racial composition cutoffs mirror Bayer et al. (2017).

The top panel of Table 1 reports the number of repeat-sales in each sample and the frequency in which the houses in the sample sold two, three, or four or more times. As noted in the previous paragraph, houses that sold once are not included in the repeat-sales sample, so the proportions are exhaustive. Regardless of the neighborhood demographics, houses that sold two times represent the majority of the repeat-sales samples. The next panel displays transaction and house characteristics for each sample. Column 2 in each panel clearly shows that the housing stock in neighborhoods that are less than 50 percent white differs from those that are greater than or equal to 50 percent white. The housing stock is generally older, smaller, and less likely to be purchased by an owner-occupier. For these reasons, among others, the average transaction price is lower.

The final panel displays the distribution of buyer race for each repeat-sales sample. Following Bayer et al. (2017), we treat 'Hispanic' as a race and use the race of both the applicant and co-applicant in the variable construction process. The results we report do not change

when we use only the primary applicant's race. The majority of the buyers in the full ($\geq$0.0) repeat-sales sample are white. However, the distribution varies according to the racial composition of the neighborhood. We examine this dynamic in greater detail in the next section.

2.1 Neighborhood Sorting

When estimating racial price differentials in housing markets one must recognize that buyers are not randomly assigned to neighborhoods. This is particularly important since the descriptive statistics in Table 1 suggest that non-majority white neighborhoods are systematically different than majority white neighborhoods. If these differences are not properly controlled for, then price estimates for black and Hispanic buyers will be biased.

To highlight the importance of controlling for neighborhood characteristics, such as racial demographics, we plot density estimates by buyer race using the percent 'Black or African American alone' 2010 census block group level estimates. Panel A of Figure 1 clearly shows that black buyers are more likely to purchase houses in neighborhoods that are majority black - especially in census block groups that are greater than 80 percent black. Whereas, non-black buyers are more likely to purchase houses in census block groups that are non-majority black. Figure 2 plots house purchases by black and white buyers on a map of Atlanta. The figure shows that black buyers are more likely to purchase houses on the south side of the city and white buyers are more likely to purchase houses on the north side of the city.

2.2 Real Estate Agent Demographics

To examine whether agent intermediation provides a mechanism through which racial price differentials can occur in housing markets we identify the race of the real estate agents involved in the transactions. Although the GAMLS data includes the name of the agents representing the buyer and seller, it does not provide demographic information about the agents. To overcome this limitation, we used the Amazon Mechanical Turk (AMT) platform

to identify the sex, race and ethnicity of the real estate agents.[7] Anonymous workers on the platform were shown images of real estate agents from www.realtor.com. As such, only agents that are National Association of Realtors (NAR) members are included in this study, which ensures that every agent received training on housing discrimination.[8] The AMT workers were shown an image of the real estate agent and asked to categorize the agent's sex as follows:

$$Sex = \begin{cases} Female \\ Male \\ No\ individual\ in\ the\ photo \\ More\ than\ one\ individual\ in\ the\ photo \end{cases}$$

If the worker selected "No individual in the photo" or "More than one individual in the photo" the agent is not included in the study. If the worker selected "Female" or "Male" then a separate set of workers categorized the agent's race and ethnicity as follows:

$$Race = \begin{cases} American\ Indian \\ Asian \\ Black\ or\ African\ American \\ Native\ Hawaiian \\ White \end{cases}$$

$$Ethnicity = \begin{cases} Hispanic \\ Non\text{-}Hispanic \end{cases}$$

[7] Additional information on Amazon Mechanical Turk is available here: https://www.mturk.com/

[8] NAR membership requires ethics training not only when agents become new members, but once every four years after. Membership also requires Realtors to pledge to conduct business in keeping with the spirit and letter of the NAR Code of Ethics. Article 10 of the Code of Ethics includes a firm statement of support for equal opportunity in housing. This ensures the agents in this study were aware of and had the knowledge and training necessary to prevent discrimination.

The race and ethnicity categories match those provided in the HMDA LAR data, allowing for ease of comparison. Each image was categorized by two independent workers. When setting up the categorization project in AMT we restricted the task to only include workers that were rated as "masters".[9] In the case of a disagreement between two workers we personally reviewed and categorized the image. We also randomly validated 1 percent of the remaining sample to ensure the categorizations are accurate.

We identify the real estate agent's race and ethnicity using an approach that is very similar to the approach employed for the LAR data. Lenders must ask the loan applicant for their sex, race and ethnicity, but cannot require that the loan applicant provide the information. If the loan application is submitted by phone, mail or internet and the applicant does not provide the information, then the information is not required for HMDA reporting purposes (FFIEC, 2013). However, if the application is submitted in person and the applicant does not provide the information, the lender is required to "note the applicant's ethnicity, race, and sex on the basis of visual observation and surname, to the extent possible."[10] Similar to the LAR data collection process, we had AMT workers classify the agents' demographic information based on a visual observation of their profile picture from the NAR website.

We merge the agent race and ethnicity classifications with the transaction data using a manually created cross reference file. We create the cross reference file to overcome two obstacles. First, there are several many-to-one and many-to-many relationships based on agent names alone. Second, agent names may change over time due to life events (e.g. marriage, divorce). We overcome these obstacles by matching agents to their unique ID in the GAMLS data based on their name *and* transaction set. Using the agent profiles on www.realtor.com we match at least one transaction listed on the agent's profile with the GAMLS data based on property address, sales price, and sale date. We then identify whether

[9] Amazon monitors worker performance over time and identifies high performing workers. Workers who demonstrate excellence across a wide range of tasks are awarded the "masters" qualification. "Masters" must continuously pass a statistical monitoring test to maintain the qualification.

[10] Additional information on the LAR data collection process available in 'Appendix B to 12 C.F.R. Part 1003' of the Guide to HMDA Reporting (FFIEC, 2013).

the agent represented the buyer or seller based on the agent's name. If there is a match we associate the sex, race, and ethnicity of the agent with their unique GAMLS ID.

AMT workers were shown 6,164 real estate agent images. 441 of the images have "No individual in the photo" or "More than one individual in the photo" and 817 images did not have a recent transaction listed on www.realtor.com. Descriptive statistics for the remaining 4,906 agents are displayed in Table 2. Approximately 70.1 percent (3,441) of the real estate agents are female and 29.9 percent (1,465) are male. White agents represent the largest racial group (79.5 percent) followed by black agents (15.2 percent), Hispanic agents (3.6 percent), and 'Asian or other' agents (1.7 percent).[11] The NAR reports that 62 percent of their members are female and that the "typical Realtor is a 53 year old white female" (NAR, 2016b), so it is no surprise that white female agents represent the majority (55.2 percent) of the sample in Table 2.

2.3 Agent Selection

When estimating racial price differentials one must consider that buyers are not randomly assigned to their real estate agent. If buyers are more likely to select an agent of the same race, then the agent's race may be used as a proxy for the buyer's race during negotiations. Table 3 displays a pairwise comparison of the race of the buyer and their agent for every matched transaction in the GAMLS dataset regardless of whether it was a repeat-sale or not. The pairwise comparison in Table 3 suggest that the race of the buyer is loosely correlated with the race of their agent. In other words, buyers disproportionately hire agents of the same race relative to the underlying real estate agent population.

'Asian or other' buyers were represented by 'Asian or other' agents in 10.1 percent of their transactions even though 'Asian or other' agents account for 1.7 percent of the agent

[11]The race of the agent is assigned using a similar process to the buyer. Any agent whose ethnicity is categorized as Hispanic is assigned to the 'Hispanic' agent group regardless of their race. Non-hispanic agents categorized as American Indian, Asian or Native Hawaiian are assigned to the 'Asian or other' agent group. Non-Hispanic black agents are assigned to the 'Black' agent group and non-Hispanic white agents are assigned to the 'White' agent group.

population. 'Asian or other' agents' 10.1 percent market share among 'Asian or other' buyers is considerably higher than black, Hispanic and white buyers. Black buyers were represented by black agents in 50.3 percent of their transactions even though black agents account for 15.2 percent of the agent population. Black buyers were the only race that did not employ white agents the majority of the time, although white agents still represented black buyers in 46.9 percent of their transactions.

Hispanic buyers were represented by Hispanic agents in 18.2 percent of their transactions even though Hispanic agents account for 3.6 percent of the agent population. Similar to 'Asian or other' buyers, Hispanic buyers hired white agents the majority of the time (73.9 percent). Similar to the other races, white buyers were disproportionately represented by agents of the same race relative to the agent population. White buyers were represented by white agents in 94.1 percent of their transactions, even though white agents account for 79.5 percent of the agent population.

Panels B and C of Figure 1 plot density estimates by race for buyer's agents and listing agents, respectively. The plots by race are similar to Panel A. Black buyer's (listing) agents are more likely to represent buyers (sellers) in neighborhoods that are majority black. Whereas, non-black buyer's (listing) agents are more likely to represent buyers (sellers) in neighborhoods that are non-majority black.

3 Racial Discrimination in Housing Markets

3.1 Fair Housing Audits

Fair housing audits are a survey technique in which minority and non-minority applicants are matched based on family and economic characteristics such that the only difference between the matched pair is, in theory, their race. The matched pairs successively visit a landlord or real estate agent in search of housing and the treatment that the two parties receive is compared. Early studies that used the fair housing audit technique find that black housing

applicants were told about 10 to 30 percent less housing units than white housing applicants (Yinger 1986; Page 1995).

Because the fair housing audits are conducted in-person their reliability is tied to how comparable the applicants are in the matched pairs. Siegelman and Heckman (1993) note that the audits are only unbiased if the applicants in the matched pairs are identical along all relevant dimensions except race. Obviously the use of human subjects violates this assumption. Recent studies have moved away from in-person audits, using online advertisements and email correspondence instead (e.g. Ahmed and Hammarstedt 2008; Hanson and Hawley 2011; Ewens et al. 2014). The switch to online correspondence has not altered the primary findings of the fair housing audits. The online correspondence studies find that landlords treat identical information from applicants with minority and non-minority sounding names differently. The in-person and online audit results establish a link between market intermediaries and differential treatment in housing markets, thereby providing motivation for the conjecture that agent intermediation may be a channel through which differential treatment occurs in housing markets.

3.2 Racial Price Differentials

Myers (2004) argues that the key to identifying racial price differentials in housing markets is to "ask whether blacks pay different amounts than whites for *identical* housing". The idiosyncratic nature of the housing market violates this assumption because no two properties are identical, and the same property may change substantially over time. Previous research attempts to overcome this obstacle using one of two distinct approaches. We briefly discuss the two approaches and then provide an overview of the methodology, including our improved approach, in the following section.

In the first approach, which we will refer to as the hedonic approach going forward, researchers isolate racial price differentials using a hedonic model that attempts to control for all characteristics of the house and surrounding neighborhood (King and Mieszkowski

1973; Yinger 1978; Chambers 1992; Kiel and Zabel 1996; Ihlanfeldt and Mayock 2009). The primary advantage of the hedonic approach is its representativeness as it includes the entire sample of housing transactions. Given the heterogenous nature of residential real estate, the primary disadvantage of the hedonic approach is its susceptibility to omitted variable bias when relevant time-invariant and time-varying characteristics are not available to the researcher.

At the house level, an omitted variable bias is an issue if the buyer's race is correlated with unobserved condition and quality. For example, if one race has a higher level of income and wealth, then members of that race will likely purchase housing that is in better condition (and/or higher quality) relative to other housing in the same neighborhood. If the condition and quality of the house are not properly controlled for, the estimated racial price differential for the wealthier race is biased upwards. Similarly, if buyers sort into neighborhoods based on their race and those neighborhoods offer different amenities that are not controlled for, then the estimated racial price differential is biased. This is a concern given the neighborhood sorting we document in Section 2.1.

In the second approach, which we will refer to as the repeat-sales approach going forward, researchers control for the time-invariant attributes of the house and neighborhood using house fixed effects (Myers 2004; Bayer et al. 2017). The primary advantage of the repeat-sales approach is its ability to compare "identical" housing when condition is held constant. However, we show that including house fixed effects alone does not yield identical housing when the condition of the house varies over time. In other words, including house fixed effects ensures the same house is compared and controls for time-invariant attributes, but it does not ensure the condition of the house remains constant across transactions.

4 Methodology

4.1 Hedonic approach

We assume the log price of house n at time t, p_{nt}, can be written as a hedonic price equation

$$p_{nt} = x_{nt}\beta + b_{nt}\tau + \mu_n + \underbrace{\psi_{nt} + v_{nt}}_{u_{nt}} \tag{1}$$

In Equation 1, x_{nt} is a vector of observable time-invariant and time-varying variables selected by the researcher, β is a vector of implicit prices, b_{nt} is either a scalar indicator variable or vector of indicator variables for the race of minority buyer, τ is the price effect associated with minority buyers' transactions, μ_n (ψ_{nt}) is a time-varying (time-invariant) effect not observed by the researcher, and v_{nt} is a zero-mean error term uncorrelated with x_{nt}, b_{nt}, μ_n, or ψ_{nt}. x_{nt} includes a set of standard house-specific attributes (square footage, bedrooms, bathrooms, etc.) and either additively or mutiplicatively separable dummy variables for the time of sale (quarter by year) and location (zip code or census tract).

We refer to μ_n and ψ_{nt} as the *quality* and *condition* of the house, respectively, and emphasize that these features are observed by the buyer, seller, and their agents, but not necessarily by the researcher. In other words, μ_n and ψ_{nt} include salient information that is not readily available to the researcher. When $\mu_n > 0$, the property is high quality and has certain time-invariant attributes that are better than average (e.g. preferable school districts or proximity to amenities). When $\psi_{nt} > 0$, the property is in great condition from either a recent capital expenditure or exceptional maintenance by the homeowner. Conversely, when $\psi_{nt} < 0$, the property is in poor condition. Indicators of properties in poor condition include property damage from fires or flood, functional obsolescence, and general neglect of necessary property maintenance.

The parameter of interest is τ. When $\tau > 0$, minority buyers pay a premium relative to non-minority buyers. When house fixed effects are not included as regressors, the least-

squares estimate of τ in Equation 1, $\hat{\tau}_H$, is biased if $\mathbb{E}[b_{nt}(\mu_n + \psi_{nt})] \neq 0$. For example, if minority buyers are more likely to purchase housing of higher (lower) quality in average condition and house-specific fixed effects are omitted, then $\mu_n > 0$ ($\mu_n < 0$) and $\hat{\tau}_H$ is upwards (downwards) biased. This observation motivates the repeat-sales approach described in the following section.

4.2 Repeat-sales approach

Using the subsample of houses that sold at least twice, the hedonic approach can easily be converted to a repeat-sales approach by including house fixed effects.[12] When house fixed effects are included as regressors, the Frisch-Waugh-Lowell theorem states that the repeat-sales estimate $\hat{\tau}_R$ is equivalent to the within-property estimator

$$p^*_{nt} = p_{nt} - \bar{p}_n = x^*_{nt}\beta + b^*_{nt}\tau + \psi^*_{nt} + v^*_{nt} \tag{2}$$

In Equation 2, p^*_{nt} is the difference between the transaction price for property n at time t and the average transaction price for house n in the data, $\bar{p}_n$. Similar notation is used for x^*_{nt}, b^*_{nt}, ψ^*_{nt}, and v^*_{nt}. By construction $\mu^*_n = 0$, so all time-invariant attributes of the house are removed regardless of whether or not they are observed by the researcher. When x_{nt} includes variables that are time-invariant or seldom change over time (square footage, bedrooms, bathrooms, etc.), $p^*_{nt} = b^*_{nt}\tau + \psi^*_{nt} + v^*_{nt}$ plus the change in local market prices over time. In equation 2, the repeat sales approach precludes μ_n from biasing $\hat{\tau}_R$. However, $\hat{\tau}_R$, remains biased upwards if $\mathbb{E}[b^*_{nt}\psi^*_{nt}] > 0$.

House fixed effects are only meaningful when house n has 2 or more transactions.[13] An

[12] The repeat-sales approach in Bailey et al. (1963), Case and Shiller (1989), and Mayer (1998) uses differenced sale prices as the dependent variable, while Bayer et al. (2017) include property-specific fixed-effects and use level sale prices as the dependent variable. Both approaches control for time-invariant attributes and are identical when each property sells exactly twice.

[13] When this is not true, the house-specific fixed effect perfectly predicts price and the transaction provides no information on β or τ.

obvious concern is that the sample of houses with repeat-sales may not be representative of the entire sample of transactions. For example, suppose a house is purchased at time t_1, renovated, and then sold shortly after at time $t_2 > t_1$. In other words, the house is flipped. For simplicity, assume these are the only two transactions for the property in the data and the renovations did not require a permit. These renovations imply

$$\mathbb{E}[\psi^*_{nt_2}] = \frac{1}{2}\mathbb{E}\left[\psi_{nt_2} - \psi_{nt_1}\right] > 0 \tag{3}$$

If white buyers purchase houses, make improvements, and then sell the renovated houses to minorities, then $\hat{\tau}_R$ is upwards biased. This occurred frequently as demonstrated by the property descriptions in Table 4. The three repeat-sales pair examples in Table 4 represent a small subsample of a large set of transactions that involve a white buyer who purchased a house, performed renovations, and then sold the house to a black buyer soon after. If not properly controlled for, the price changes associated with the renovations bias $\hat{\tau}_R$ upwards and $\hat{\tau}_R > 0$ can be misinterpreted as a black-white price differential.

Including regressors that control for major renovations can mitigate bias, but such controls are not available in most datasets. For example, county assessor offices frequently record time-invariant, house-specific attributes but rarely record time-varying, house-specific attributes. Sufficiently large capital improvements that require a building permit are recorded, but may not be available in transaction databases maintained by the county. Moreover, modest capital improvements including new cabinetry, new appliances, and other cosmetic improvements that do not require a building permit are not available in transaction databases maintained by the county or municipality.

One way to remove the bias associated with flipped properties is to exclude repeat-sales with short holding periods from the sample. After running baseline estimates with the full repeat-sales sample we exclude repeat-sales with holding periods of three years or less. Holding period filters are commonly used in the real estate literature. For example, Levitt

and Syverson (2008) drop any house that is sold twice within a three year period "due to concerns that the house has been purchased and rehabbed for sale". Of course, there is no guarantee the properties that remain after filtering are representative of the entire sample of transactions.

Alternatively, Liu et al. (2018) describe methods for incorporating textual information about the house in a hedonic pricing model in order to mitigate an omitted variable bias that does not require excluding transactions from the data set. The MLS remarks section (*remarks*) provides an unstructured written description of the house provided by the listing agent at the time the house is listed for sale. As indicated in Table 4, the remarks include time-varying information about the house that is not necessarily available in databases maintained at the county level. In the following subsections, we summarize the methodology in Liu et al. (2018) and highlight the extension we employ in this paper to identify and delineate between the time-invariant and time-varying attributes of the house.

4.3 Textual analysis

Nowak and Smith (2017) and Liu et al. (2018) show that the textual information in the MLS remarks can be used to proxy for both μ_n and ψ_{nt}. A *token* refers to any word or phrase in the remarks and the set of the $K = 2,000$ most frequent tokens form a candidate token set. The candidate tokens are chosen based on their frequency in the remarks and not on any ex-ante beliefs about their ability to proxy for $\mu_n + \psi_{nt}$. In this way, we remain agnostic as to which tokens represent the condition and quality of the property. The information in the tokens can be incorporated into a hedonic model using indicator variables where $w_{ntk} = 1$ if token k is in the agent's remark for house n at time t and $w_{ntk} = 0$, otherwise.

Text, by nature, is high-dimensional. As such, the number of indicator variables can be large and parameter estimates can overfit the data for even moderately large data sets. However, we assume that $Q \ll K$ *relevant tokens* can provide a sufficient approximation to $\mu_n + \psi_{nt}$ such that

$$\mu_n + \psi_{nt} = \sum_{k \in S} w_{ntk}\theta_k + r_{nt} \tag{4}$$

In Equation 4, $S \subset \{1, ..., K\}$ is the index of the Q relevant tokens, θ_k is the implicit price of token k, and r_{nt} is an approximation error. When the approximation error is uncorrelated with b_{nt}, S is known, the indicator variables associated with S are included alongside x_{nt} and b_{nt} in a regression, and the least-squares estimate of τ is consistent and has an asymptotically normal distribution. In practice, S is unknown and this estimation procedure in infeasible.

However, a feasible estimator using an estimate of S is possible using double-selection methods that solve the following[14]

$$(\hat{\beta}_p', \hat{\tau}_p, \hat{\theta}_p')' = \arg\min_{\beta,\tau,\theta} \sum (p_{nt} - x_{nt}\beta - b_{nt}\tau - w_{nt}\theta)^2 + \lambda \sum_k |\theta_k \phi_{p,k}| \tag{5}$$

$$(\hat{\beta}_b', \hat{\theta}_b')' = \arg\min_{\beta,\theta} \sum (b_{nt} - x_{nt}\beta - w_{nt}\theta)^2 + \lambda \sum_k |\theta_k \phi_{b,k}| \tag{6}$$

$\lambda, \phi_{p,k}$, and $\phi_{d,k}$ are positive scalars that penalize θ_k.[15] Equation 5 is an ℓ_1 penalized regression and a variant of the LASSO.[16] The shape of the penalty performs variable selection by setting some parameters in $\hat{\theta}_p$ and $\hat{\theta}_b$ to 0. The loadings ϕ_{pk} and ϕ_{bk} control for heteroscedasticity in the error term. The index of the $\hat{Q}_p$ non-zero coefficients in $\hat{\theta}_p$ is defined as $\hat{S}_p \subset \{1, ..., K\}$ and similarly for $\hat{Q}_b$ and $\hat{S}_b$.

Intuitively, solving Equation 5 is analogous to finding the set of tokens, both time-invariant and time-varying, that best predict price. However, the set of variables that are

[14] We summarize the variable selection procedure here and refer the reader to Belloni et al. (2014) and Liu et al. (2018) for a more in-depth discussion.

[15] $\lambda = 2c\sqrt{N}\Phi^{-1}(1 - \gamma/2K)$ where N is the total number of transactions, $c = 1.10$, and $\gamma = 0.10$. The choices of c, γ are recommended in Belloni et al. (2014). We show the choice of penalty parameters has a neglible effect on $\hat{\tau}_R$ in an internet appendix. The infeasible estimator in Equation 5 uses the penalty parameters $\phi_{p,k} = \sqrt{N^{-1}\sum_n w_{ntk}^2 v_{nt}^2}$ where v_{nt} is the error term in the hedonic equation. Feasible choices for $\phi_{p,k}, \phi_{b,k}$ are determined using an iterative procedure described in Belloni et al. (2014).

[16] LASSO: Least Absolute Shrinkage and Selection Operator

necessary for resolving the omitted variable bias may include weak predictors of price not included in $\hat{S}_p$. However, solving Equation 6 mitigates this concern by identifying strong predictors of b_{nt}. Define $\hat{S}_2 = \hat{S}_p \cup \hat{S}_b$ as the index of tokens with non-zero coefficients in either $\hat{S}_p$ or $\hat{S}_b$. Tokens in $\hat{S}_2$ are either strong predictors of price or moderate predictors of price that are strong predictors of b_{nt}. Alternatively, the omission of tokens not in $\hat{S}_2$ from a regression is unlikely to yield biased estimates of τ.

Belloni et al. (2014) define the *post double-selection* estimator as the least-squares estimator $\hat{\tau}_2$ using x_{nt} and tokens in $\hat{S}_2$ as controls. Of course, when $\hat{S} \neq S$, variable selection mistakes have occurred, and the model is misspecified. However, when a sparse θ_p provides a reasonable approximation to $\mu_n + \psi_{nt}$, the selection criteria in Equations 5 and 6 ensure variables in S excluded from $\hat{S}$ have a negligible impact on $\hat{\tau}_2$. Moreover, the post double-selection estimator is consistent and asymptotically normally distributed (Belloni et al., 2014).[17] Moreover, for our purposes, the set of tokens in $\hat{S}$ is sufficient for removing the omitted variable bias in the estimate of τ associated with $\mu_n + \psi_{nt}$.

4.4 Time-varying tokens

The previous section describes a method to identify relevant time-invariant and time-varying tokens. This section extends the variable selection procedure in Liu et al. (2018) to show that a repeat-sales method can be used to identify the subset of tokens in S that proxy for the time-varying attributes in ψ_{nt}. Using the same set of $K = 2,000$ candidate tokens and repeat-sales transaction sample, we can identify relevant time-varying tokens by solving

$$(\hat{\beta}_p^{*\prime}, \hat{\tau}_p^*, \hat{\theta}_p^{*\prime})' = \arg\min_{\beta,\tau,\theta} \sum (p_{nt}^* - x_{nt}^*\beta - b_{nt}^*\tau - w_{nt}^*\theta)^2 + \lambda_p \sum_k |\theta_k \phi_{p,k}^*| \tag{7}$$

$$(\hat{\beta}_b^{*\prime}, \hat{\theta}_b^{*\prime})' = \arg\min_{\beta,\theta} \sum (b_{nt}^* - x_{nt}^*\beta - w_{nt}^*\theta)^2 + \lambda_b \sum_k |\theta_k \phi_{b,k}^*| \tag{8}$$

[17] See Theorem 1 of Belloni et al. (2014) for details on the assumptions and Corollaries 2 and 3 for linear models.

Where the token indicators are demeaned at the property level as $w^*_{ntk} = w_{ntk}$, $\hat{S}^*_p$ and $\hat{S}^*_b$ is the index of the non-zero coefficients in $\hat{\theta}^*_p$ and $\hat{\theta}^*_b$, respectively, and $\hat{S}^*_2 = \hat{S}^*_p \cup \hat{S}^*_b$. Solving Equations 7 and 8 is identical to Equations 5 and 6 using property demeaned variables. However, tokens in $\hat{S}^*_p$ indicate relevant tokens after controlling for time-invariant factors at the property level. Alternatively,

$$\psi_{nt} = \sum_{k \in \hat{S}^*_p} w^*_{nt} \theta^*_{pk} + r^*_{nt} \tag{9}$$

In Equation 9, r^*_{nt} is an approximation error. Although tokens that indicate time-invariant attributes of the property may be included in the candidate set, these tokens will not have any impact on p^*_{nt} and, with perfect model selection, will not be included in $\hat{S}^*_2$.

The enhanced variable selection procedure identifies the subset of relevant tokens that approximate the "unobserved" time-varying attributes (ψ^*_{nt}) in Equation 2. By construction, the complementary subset of tokens in $\hat{S}_2$ that are not in $\hat{S}^*_2$ correspond with the "unobserved" time-invariant attributes (μ_n) in Equation 1. When running the empirical analysis, we include the time-varying and time-invariant tokens separately in Equation 2 to demonstrate the efficacy of our approach. In doing so, we show that it is (is not) the time-varying (time-invariant) attributes of the house that bias the racial price differentials in the extant literature.

Belloni et al. (2016) describe the cluster-LASSO that accounts for the panel structure of data when selecting $\hat{S}^*_2$. As pointed out in Belloni et al. (2016), ignoring within property correlation induced by the within transformation may lead to ϕ^*_{pk} and ϕ^*_{bk} that are too small and a set of relevant tokens that is too large. If the additional tokens were selected at random, this would not normally distort estimates and standard errors. However, the additional tokens selected are the most correlated with the noise and can have a significant impact on statistical inference. Theorem 3 in Belloni et al. (2016) demonstrates that when penalty loadings in ϕ^*_{pk} and ϕ^*_{bk} account for the panel structure of the data, tokens in $\hat{S}^*_2$

can lead to valid statistical inference on τ. Where applicable, we provide results for both the heteroscedastic-LASSO in Belloni et al. (2014) and the cluster-LASSO in Belloni et al. (2016).

5 Results

5.1 Baseline hedonic and repeat-sales

The estimates in Table 5 provide a baseline for comparing the approach we employ in subsequent analysis. Panel A of Table 5 uses the hedonic approach and Panel B uses the repeat-sales approach. Similar to the extant literature, the two approaches yield conflicting results. The hedonic approach estimates that black buyers pay, on average, 1.9 percent less than white buyers. While the repeat-sales approach estimates that black buyers pay, on average, 3.6 percent more than white buyers. The disparity is likely related to (i) black buyers sorting into different neighborhoods/housing that are not properly controlled for in the hedonic approach ($\mathbb{E}[b_{nt}(\mu_n + \psi_{nt})] < 0$) and (ii) improvements to time-varying attributes that are not properly controlled for in the repeat-sales approach ($\mathbb{E}[\psi_{nt}] > 0$). We address these issues using a repeat-sales approach that either employs a series of filters or includes time-varying tokens that are selected using our enhanced textual analysis procedure.

5.2 Repeat-sales with filters

Table 6 reports black-white and Hispanic-white price differentials using the repeat-sales approach. The racial price differentials are organized in two panels. Panel A presents the black-white price differentials and Panel B presents the Hispanic-white price differentials for the entire repeat-sales sample and several filtered subsamples. Both panels report estimates relative to non-Hispanic white buyers.

Four distinct specifications that differ only in their underlying neighborhood racial composition are reported in columns 1 to 4 of both panels. Column 1 includes all repeat-sales

transactions (percent white $\geq$ 0.0). Column 2 includes repeat-sales transactions in neighborhoods that are less than 50 percent white. Column 3 includes repeat-sales transactions in neighborhoods that are greater than or equal to 50 percent white and column 4 includes transactions in neighborhoods that are greater than or equal to 80 percent white. Unsurprisingly, the proportion of black-white transactions in column 2 is considerably greater than column 4.[18]

The baseline black-white price differentials reported in Panel A of Table 6 match the repeat-sales estimates in Panel B of Table 5. The baseline black-white price differential is 3.6 percent in column 1 of Panel A. However, the black-white coefficient estimates vary considerably by neighborhood racial composition in columns 2 to 4. The Hispanic-white price differential is 1.2 percent for the full repeat-sales sample in Panel B. However, the Hispanic-white estimates are statistically insignificant in columns 2 to 4 when the neighborhoods are partitioned based on their racial composition.

We apply a series of cumulative filters to ensure the repeat-sales sample is relatively homogeneous and remove repeat-sales pairs that are most susceptible to a time-varying omitted attribute bias. The first filter uses the occupancy field in the LAR data to identify and remove transactions that are not "owner-occupied as a principal dwelling". The filter removes rental properties and houses that were purchased by investors. We remove rental houses since they are typically of lower quality and have more wear and tear relative to owner-occupied houses (Wang et al., 1991). The owner-occupied filter reduces the magnitude of both the black-white and Hispanic-white price differentials. Additionally, the Hispanic-white price differential is no longer statistically significant.

The next set of filters address the concern that houses purchased by buyers of different races may have undergone differential amounts of renovation or maintenance during the holding period of the previous owner. We filter out houses that were flagged in the tax assessor

[18]We provide a detailed breakdown of the transaction type by race for each subsample in the appendix. We partition the neighborhoods by percent white to facilitate comparison with the estimates in Bayer et al. (2017). In an internet appendix we provide estimates using percent black to partition the neighborhood racial composition in columns 2 to 4. The estimates are similar for the filtered subsamples.

data as having undergone a major renovation and transactions with a holding period of three years or less. The holding period filter removes houses that were purchased, renovated, and resold in a short period of time (i.e. flipped), thereby limiting the effect of the renovations that were performed on the racial price differential estimates. After applying the second set of filters, the black-white and Hispanic-white price differentials are no longer statistically significant.

The final filter removes all repeat-sales transactions in which the house was involved in at least one distressed (shortsale or REO) transaction during the study period. The filters are cumulative, so the final row of Table 6 includes repeat-sales transactions that are: owner-occupied, not flipped or remodeled, and not distressed. The statistically insignificant racial price differentials show that minority and non-minority buyers pay a similar price for comparable housing.

5.3 Repeat-sales with tokens

The filtered results in Table 6 suggest the baseline estimates of $\hat{\tau}_R$ are biased by "unobserved" changes to the time-varying attributes of the house. To further investigate this claim, we augment the repeat-sales regression with information about the property that is easily observed and available, but frequently excluded since it is provided to the researcher in a high-dimensional text format. More specifically, we include indicator variables for key words and phrases (tokens) from the agents' remarks about the property. The remarks field is only available in the GAMLS data, so the racial price differentials in this section are estimated using the repeat-sales sample in Panel B of Table 1.

The three panels in Table 7 incorporate three different types of tokens. Panel A incorporates unigram (one word) tokens, Panel B incorporates bigram (two word) tokens, and Panel C incorporates flex-gram (multi-word) tokens. Additional information about the token creation process is available in the appendix. Column 1 of each panel provides a baseline result prior to incorporating the tokens. The remaining columns present specificiations that

incorporate either time-varying tokens (column 2), time-invariant tokens (column 3), or both (column 4).

Prior to introducing the tokens, the black-white price differential in column 1 of Panel A is 1.9 percent. However, after we incorporate the time-varying tokens in column 2, the black-white price differential decreases to 0.9 percent and is no longer statistically significant. This result provides additional evidence that the racial price differentials reported in the extant literature are biased by changes to the time-varying attributes of the house that are not properly controlled for in the regression.

Figure 3 displays ten of the most positive (e.g. *renovated, new,* and *less*) and negative (e.g. *shortsale, asis,* and *opportunity*) time-varying tokens. At first glance some of the tokens may seem contradictory. For example, the word *less* typically takes on a negative connotation and the word *opportunity* typically takes on a positive connotation. However, houses that have a "roof that is *less* than two years old" sell for a premium and houses that offer an "*opportunity*" often need some work and sell for a discount. The use of bigrams and more elaborate flex-grams help address this confusion but do not affect estimates of τ. More importantly, the tokens in $\hat{S}_2^*$ represent time-varying attributes of the house that, when not properly controlled for, bias the racial price differentials in column 1.

The third column of Panel A incorporates only the time-invariant unigram tokens. In contrast to column 2, the black-white price differential is statistically significant at the 5 percent level. This result bolsters the claim that the tokens in column 2 control for the time-varying attributes of the house that bias black-white price differentials in Column 1 and in the extant literature. The time-invariant unigram tokens identify positive and negative attributes of the house that do not change over time. For example, several local neighborhoods are included in both the positive (e.g. *brookhaven, oakhurst,* and *roswell*) and negative (e.g. *parkview and kennesaw*) tokens. Given that the repeat-sales approach differences out time-invariant attributes of the house, such as the neighborhood it is located in, it is no surprise that including these attributes in the repeat-sales specification has a neglible effect on the

racial price differential coefficient estimates.

Column 4 incorporates both time-varying and time-invariant tokens. Similar to column 2, the black-white price differential is statistically insignificant. Overall, the enhanced variable selection procedure does a good job of delineating between the time-varying and time-invariant tokens in the remarks. There are, however, a few time-invariant tokens that one would expect, ex-ante, to be included in the time-varying tokens (and vice-versa). For example, the *vacant* token is classified as time-invariant in Figure 3 although it has time-varying features. However, houses that are vacant when they are listed for sale at time t_1 may have a higher probability of being vacant the next time they are put up for sale at time t_2. In which case, it makes sense that the *vacant* token is classified as time-invariant.[19]

The results in Panels B and C of Table 7 are nearly identical to Panel A. Regardless of the type of token employed (unigram, bigram, or flex-gram) the racial price differential is statistically insignificant when the time-varying textual information is included. We provide a description and example of every time-varying and time-invariant unigram token in Figure 3 in an internet appendix. Additional insight into the bigram and flex-gram tokens is also provided in an internet appendix.

5.4 Finite Sample Considerations

The post double-selection estimator is an asymptotic result. To establish the finite sample credibility of our claims that the repeat-sales post double-selection estimate in the previous subsection is statistically insignificant, we perform a simulation experiment. The simulations determine the power of the post double-selection estimator to detect $\tau \in \{0.005, 0.010, ..., 0.020\}$. The τ considered correspond to previously reported estimates in the literature as well estimates in Table 7. The variables p_{nt} and b_{nt} are generated using parameters estimated from the data for 500 simulations of each τ. Further details of the simulation experiment are

[19]For example, it is doubtful that a "flipper" will buy a house if it is occupied by a tenant with a long-term lease. Instead, they are more likely to target vacant fixer-uppers that they start renovating immediately. In which case, the house is vacant when they purchase it and when they sell it.

described in an internet appendix.

Results of the simulations are displayed in Figure 4. The two lines correspond to the fraction of simulations where $H_0 : \tau = 0$ is rejected using either a 5% or 10% significance level, respectively. Standard errors are heteroscedastic consistent. Standard errors clustered at the property level yield similar results and are available upon request. The results of the simulation suggest that the procedure described in the paper can reliable detect $\tau \geq 0.01$. This result is encouraging given previous point estimates in the literature suggest $\tau \approx 0.02$.

5.5 Financing controls

The results presented in the preceding sections provide strong statistical evidence that minority and non-minority buyers pay a similar price for comparable housing. The estimates, however, could be correlated with other buyer attributes that affect the home buying process. Table 8 presents results from additional specifications that control for the buyer's access to financing and financial position. This is particularly important given that Ondrich et al. (2003) find evidence of statistical discrimination that they attribute to an agent's uncertainty about black buyers' ability to put forth successful bids.

Prior to representing a buyer in the home search process most real estate agents require the potential buyer to be either prequalified or preapproved for a loan. The prequalification process relies on the loan applicant's self-reported assets, debt, income, and credit score. Based on this self-reported information, the lender provides an estimate of the loan amount that the applicant qualifies for. However, the lender provides no guarantee that they will approve the loan in the prequalification letter. In contrast, the preapproval process provides "a written commitment to the applicant valid for a designated period of time to extend a home purchase loan up to a specified amount."[20] Thus, the preapproval process should negate the agent uncertainty in Ondrich et al. (2003).

The first column of Table 8 provides baseline estimates similar to the racial price dif-

[20]See Section 203.2(b)(2) of Regulation C in the Home Mortgage Disclosure Act for more information.

ferentials in the final row of column 1 in Table 6. Note, however, that the repeat-sales samples in Tables 6 and 8 differ slightly since we drop repeat-sales pairs in Table 8 where the income of the buyer is not available for at least one of the transactions. In column 2 we include an indicator variable that identifies whether the buyer was preapproved by their lender. The black-white coefficients do not change when the preapproval variable is included in the specification. The coefficient on the preapproval variable is negative, which suggests that getting preapproved may reduce the sales price, although the coefficient is statistically insignificant.[21]

Columns 3 and 4 report results for specifications that add the buyer's income and down payment percentage. Bayer et al. (2017) note that these variables may be correlated with sales price for a number of reasons, including the ability to secure financing and differences in search costs. The inclusion of these financial controls has a neglible impact on the racial price differentials, thereby providing additional evidence that minorities and non-minorities pay a similar price for comparable housing.

5.6 Agent racial price differentials

Agent intermediation should, in theory, eliminate racial price differentials in housing markets. If the seller and their listing agent never meet the buyer, then the buyer's race is not revealed. In which case, sellers cannot discriminate against the buyer. The racial price differentials we present in the preceding sections support this conjecture. However, the fact that buyers sort into agent representation based on race (see Section 2.3) suggests that the race of buyer's agent may serve as a proxy for the race of the buyer. As such, we test for racial price differentials at the agent level. More specifically, we test whether buyers represented by black agents pay a similar price as buyers represented by white agents for comparable housing.

[21] Approximately 4.7 percent of the repeat-sales transactions involved a buyer that was preapproved. 4,863 of the repeat-sales transactions involved black buyers. Of which, approximately 4.1 percent were preapproved. Interacting the preapproval and race variables in the specification did not affect the results. The interaction term was also statistically insignificant.

Table 9 presents black-white price differentials for the repeat-sales transactions in which we know the race of the buyer's agent.[22] The format of Table 9 is similar to Table 6 in that we (i) report estimates for four distinct specifications that differ only in the underlying neighborhood racial composition and (ii) apply a series of filters to remove repeat-sales transactions that are most susceptible to a time-varying attribute bias. The baseline black-white price differential is statistically insignificant in all but one column of the first row. This finding holds as we filter the data to remove non-owner-occupied purchases, short holding periods, and distressed transactions. Thus, we conclude that buyers represented by black agents and white agents pay a similar price for comparable housing.

5.7 Study Limitations

Although the dataset utilized in this study is rich in terms of observable house, neighborhood, and buyer attributes it does have several limitations. The collection and classification of the real estate agents' demographic information was performed once, so the racial profile of the agent network is static. In reality, the agent network is dynamic so its racial profile likely changes over time.[23] Because the agent network in this study represents a snapshot in time we recognize that there may be a survivorship bias in terms of the agents included in this study. However, given the time consuming task and difficulty/cost of identifying the race of agents who are no longer active in the network we leave this inquiry for future research.

The inability to consistently identify agent race in historical transactions also limits our ability to run a repeat-sales specification that includes the race of both the buyer and their agent. Ideally, we would test whether buyers represented by agents of a different race pay a similar price relative to buyers represented by agents of the same race. For example, we would test whether black buyers represented by white agents pay a similar price as black buyers represented by black agents. Unfortunately, this analysis is not possible due to the

[22] Descriptive statistics for the buyer's agent repeat-sales sample are provided in the appendix.

[23] This is especially true in the late-2000s when the economy went into a recession. During this time period, membership in NAR dropped from a high of 1,357,732 in 2006 to 999,824 in 2012 according to the NAR's historical data on membership (NAR, 2017).

limited number of repeat-sales in which we know the race of both the buyer and their agent.

We also recognize that membership in the National Association of Realtors (NAR) is optional, and that we were forced to drop 1,258 of the 6,164 agent records listed on the NAR website because they were missing a photo or did not have a matching transaction in the GAMLS data. If membership in NAR or posting a photo on www.realtor.com is correlated with agent race, then the racial profile of the agent network we present may be skewed. To partially address this concern we compared the number of agents in this study (4,906) to the number of unique agent IDs in the GAMLS dataset with at least one transaction in 2016. The results show that the agent network we examine represents a large proportion (68.7 percent) of the agents that had at least one transaction in 2016. In addition to its representativeness, we chose to identify the race of agents using photos from the NAR website because it ensures that the agents included in our study received training on housing discrimination and fair housing policy.

6 Conclusion

Whether minorities pay more than non-minorities for comparable housing has important policy implications. Thus, it is no surprise that it remains a central concern in the literature studying racial discrimination in housing markets. After replicating the racial price differentials reported in the extant literature we show that the estimates suffer from an omitted variable bias stemming from the time-varying attributes of the house. We control for the time-varying attributes using two distinct approaches. The first approach employs a series of filters to remove repeat-sales transactions that are most susceptible to the time-varying attribute bias. The second approach controls for the time-varying attributes using textual analysis instead of filters. Regardless of the approach taken we find that minority and non-minority buyers pay a similar price for comparable housing.

The textual analysis we employ highlights the fact that the repeat-sales approach does

not (does) control for the time-varying (time-invariant) attributes of the house. Using a novel variable selection procedure, we carefully delineate between time-invariant and time-varying tokens (i.e. textual information) and then include the tokens separately in the repeat-sales specification. As expected, the inclusion of the time-invariant tokens has no effect on the racial price differentials. However, the inclusion of the time-varying tokens renders the racial price differentials statistically insignificant.

We also examine the dimensions along which differential treatment can occur in housing markets. We show that buyers are disproportionately represented by agents of the same race relative to the underlying real estate agent population. As a consequence, the agent's race may serve as a proxy for the buyer's race during negotiations. Using a repeat-sales estimation that incorporates the race of the buyer's agent, we find no evidence of racial price differentials at the agent level. Although we find no evidence of price-based racial discrimination at the buyer or agent level, the results we report do not rule out other forms of racial discrimination in housing markets (steering, blockbusting, etc.).

References

Ahmed, A. M. and Hammarstedt, M. (2008). Discrimination in the rental housing market: A field experiment on the internet. *Journal of Urban Economics*, 64(2):362–372.

Bailey, M. J., Muth, R. F., and Nourse, H. O. (1963). A regression method for real estate price index construction. *Journal of the American Statistical Association*, 58(304):933–942.

Baker, S. R., Bloom, N., and Davis, S. J. (2016). Measuring economic policy uncertainty. *The Quarterly Journal of Economics*, 131(4):1593–1636.

Bayer, P., Casey, M. D., Ferreira, F., and McMillan, R. (2017). Racial and ethnic price differentials in the housing market. *Journal of Urban Economics*, 102:91–105.

Belloni, A. and Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650.

Belloni, A., Chernozhukov, V., Hansen, C., and Kozbur, D. (2016). Inference in high-dimensional panel models with an application to gun control. *Journal of Business & Economic Statistics*, 34(4):590–605.

Case, K. E. and Shiller, R. J. (1989). The efficiency of the market for single-family homes. *The American Economic Review*, 79(1):125–137.

Chambers, D. N. (1992). The racial housing price differential and racially transitional neighborhoods. *Journal of Urban Economics*, 32(2):214–232.

Ewens, M., Tomlin, B., and Wang, L. C. (2014). Statistical discrimination or prejudice? a large sample field experiment. *Review of Economics and Statistics*, 96(1):119–134.

FFIEC (2013). A Guide to HMDA Reporting - Getting it Right!

Hanson, A. and Hawley, Z. (2011). Do landlords discriminate in the rental housing market? evidence from an internet field experiment in us cities. *Journal of Urban Economics*, 70(2):99–114.

Ihlanfeldt, K. and Mayock, T. (2009). Price discrimination in the housing market. *Journal of Urban Economics*, 66(2):125–140.

Kiel, K. A. and Zabel, J. E. (1996). House Price Differentials in U.S. Cities: Household and Neighborhood Racial Effects. *Journal of Housing Economics*, 5(2):143–165.

King, A. T. and Mieszkowski, P. (1973). Racial Discrimination, Segregation, and the Price of Housing. *Journal of Political Economy*, 81(3):590–606.

Levitt, S. D. and Syverson, C. (2008). Market Distortions When Agents Are Better Informed: The Value of Information in Real Estate Transactions. *The Review of Economics and Statistics*, 90(4):599–611.

Liu, C., Nowak, A., and Smith, P. (2018). Asymmetric or incomplete information about asset values? *Working Paper*.

Loughran, T. and Mcdonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, 66(1):35–65.

Mayer, C. J. (1998). Assessing the performance of real estate auctions. *Real Estate Economics*, 26(1):41–66.

Myers, C. K. (2004). Discrimination and neighborhood effects: understanding racial differentials in US housing prices. *Journal of Urban Economics*, 56(2):279–302.

NAR (2016a). National association of realtors 2016 profile of home buyers and sellers.

NAR (2016b). National association of realtors member profile.

NAR (2017). National association of realtors historic membership report.

Nowak, A. and Smith, P. (2017). Textual Analysis in Real Estate. *Journal of Applied Econometrics*, 32(4):896–918.

Ondrich, J., Ross, S., and Yinger, J. (2003). Now you see it, now you don't: why do real estate agents withhold available houses from black customers? *Review of Economics and Statistics*, 85(4):854–873.

Page, M. (1995). Racial and ethnic discrimination in urban housing markets: Evidence from a recent audit study. *Journal of Urban Economics*, 38(2):183–206.

Scharnhorst, E. (2017). Nearly half of minority homebuyers in a 2016 housing market survey felt they may have been discriminated against when trying to buy a home. https://www.redfin.com/blog/2017/02/minority_homebuyers_market_survey.html. [Online; accessed 09-13-2017].

Siegelman, P. and Heckman, J. (1993). The urban institute audit studies: Their methods and findings. In Fix, M. and Struyk, R. J., editors, *Clear and Convincing Evidence: Measurement of Discrimination in America*, chapter 5, pages 187–258. Urban Institute Press, Lanham, MD.

Wang, K., Grissom, T. V., Webb, J. R., and Spellman, L. (1991). The impact of rental properties on the value of single-family residences. *Journal of Urban Economics*, 30(2):152–166.

Yinger, J. (1978). The Black-White Price Differential in Housing: Some Further Evidence. *Land Economics*, 54(2):187–206.

Yinger, J. (1986). Measuring Racial Discrimination with Fair Housing Audits: Caught in the Act. *The American Economic Review*, 76(5):881–893.

Yinger, J. (1995). *Closed Doors, Opportunities Lost: The Continuing Costs of Housing Discrimination*. Russell Sage Foundation.

Table 1: Descriptive statistics for the repeat-sales transaction data

	Panel A: CoreLogic Buyer				Panel B: GAMLS Buyer			
	≥0.0 (1)	<0.5 (2)	≥0.5 (3)	≥0.8 (4)	≥0.0 (1)	<0.5 (2)	≥0.5 (3)	≥0.8 (4)
Observations	97,001	32,013	64,988	27,619	35,224	10,712	24,512	9,621
Repeat-sales (proportions)								
Sold twice	0.80	0.81	0.79	0.79	0.88	0.90	0.87	0.86
Sold three times	0.17	0.16	0.18	0.19	0.11	0.09	0.12	0.13
Sold four or more times	0.03	0.03	0.03	0.03	0.01	0.01	0.01	0.01
House characteristics								
Price (000s)	256.28	167.80	299.87	363.08	242.46	165.65	276.02	332.77
Sfla (000s)	2.29	1.85	2.51	2.70	2.29	1.95	2.44	2.60
Age	28.28	31.47	26.72	30.04	26.88	27.36	26.66	30.92
Bedrooms	3.41	3.18	3.52	3.61	3.69	3.48	3.78	3.89
Bathrooms	2.72	2.32	2.92	3.08	2.61	2.35	2.72	2.86
Owner-occupier	0.90	0.83	0.93	0.94	0.93	0.90	0.94	0.95
Buyer race (proportions)								
Asian and other	0.09	0.10	0.09	0.07	0.09	0.10	0.08	0.06
Black	0.22	0.48	0.09	0.05	0.19	0.43	0.09	0.05
Hispanic	0.06	0.08	0.06	0.04	0.07	0.08	0.06	0.04
White	0.63	0.34	0.77	0.85	0.65	0.38	0.77	0.85

Notes: This table provides summary statistics for the repeat-sales data used in the empirical analysis. The data includes single-family detached houses in Atlanta, Georgia that transacted more than once from January 2000 through September 2016. Panel A includes every repeat-sales transaction in which we successfully match a transaction in the CoreLogic data with a loan in the LAR data. Panel B is a subsample of Panel A in which we successfully merged the CoreLogic-LAR data with the GAMLS data. Columns 2 to 4 are filtered by neighborhood racial composition in both panels. Column 2 includes repeat-sales transactions in neighborhoods that are less than 50% white. Column 3 (4) includes repeat-sales transactions in neighborhoods that are greater than or equal to 50% (80%) white.

Table 2: Agent race and sex

	Female (1)	Male (2)	Total (3)
White	2,709	1,192	3,901
	(55.2%)	(24.3%)	(79.5%)
Black	546	202	748
	(11.1%)	(4.1%)	(15.2%)
Hispanic	129	46	175
	(2.6%)	(0.9%)	(3.6%)
Asian and other	57	25	82
	(1.2%)	(0.5%)	(1.7%)
Total	3,441	1,465	4,906
	(70.1%)	(29.9%)	(100.0%)

Notes: This table cross tabulates the race and sex of the 4,906 real estate agents included in this study. The agents' race is tabulated vertically and their sex is tabulated horizontally in descending order.

Table 3: Buyer and agent race

	Buyer's Race				
	Asian	Black	Hispanic	White	Total
	(1)	(2)	(3)	(4)	(5)
Asian and other	228	55	22	234	539
	(10.1%)	(1.1%)	(1.3%)	(0.9%)	(1.5%)
Black	159	2,579	115	630	3,483
	(7.0%)	(50.3%)	(6.6%)	(2.4%)	(9.9%)
Hispanic	143	92	317	678	1,230
	(6.3%)	(1.8%)	(18.2%)	(2.6%)	(3.5%)
White	1,726	2,406	1,285	24,638	30,055
	(76.5%)	(46.9%)	(73.9%)	(94.1%)	(85.1%)
Total	2,256	5,132	1,739	26,180	35,307
	(6.4%)	(14.5%)	(4.9%)	(74.1%)	

Notes: This table displays the total number of transactions by the race of the buyer and their agent. The number of transactions by buyer race is tabulated vertically and the number of transactions by agent race is tabulated horizontally. The buyer and their agent's race are both sorted alphabetically.

Table 4: Repeat-sales with short holding periods

Repeat-sales pair #1			
Zip code	Sale date	Price	Buyer race
30144	3/31/2005	$107,000	White
MLS Remark: fixer upper on a great street in kennesaw! 4 sides brick hardwoods throughout, huge kitchen. good bones. renovate and sell or renovate and rent. sold as is.			
—"—	6/12/2006	$155,000	Black
MLS Remark: fantastic like new home in historic kennesaw. Completely renovated with new roof, gutters, hardwoods, paint, cabinets, baths, appliances, and fenced back yard.			

Repeat-sales pair #2			
Zip code	Sale date	Price	Buyer race
30084	1/3/2006	$175,000	White
MLS Remark: fabulous brick home. open floor plan, oversized rooms, large master. large landscaped lot. quick access to 285, northlake mall. needs some tlc. call for offer details.			
—"—	7/27/2006	$239,000	Black
MLS Remark: newly renovated blonde brick ranch on large landscaped lot. updated kitchen and baths w/new corian & marble countertops. new lighting, gleaming hardwoods & tile fireplaces. must see all the updates!			

Repeat-sales pair #3			
Zip code	Sale date	Price	Buyer race
30039	8/3/2005	$105,000	White
MLS Remark: two story brick in norris lake, bank foreclosure needs tlc.			
—"—	12/22/2005	$215,000	Black
MLS Remark: a must see, new carpet, new paint, new garage door w/ opener, 96 acre lake community w/playground, pool, tennis, & fishing. relandscaped yard, new septic tank, completely renovated.			

Note: This table displays three repeat-sales transaction pairs in which the first transaction involved a white buyer and the second transaction involved a black buyer. The repeat-sales pairs were selected to provide examples of black-white transactions that had short holding periods.

Table 5: Baseline black-white price differentials

	Panel A: Hedonic				Panel B: Repeat-sales			
	≥0.0	<0.5	≥0.5	≥0.8	≥0.0	<0.5	≥0.5	≥0.8
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
Black buyer	−0.019***	−0.013***	−0.031***	−0.041***	0.036***	0.059***	0.022***	0.021
	(0.002)	(0.002)	(0.002)	(0.004)	(0.006)	(0.016)	(0.006)	(0.015)
N	160,844	59,753	101,091	39,453	97,001	32,013	64,988	27,619
R^2	0.876	0.770	0.884	0.877	0.985	0.971	0.987	0.989
Controls	✓	✓	✓	✓	✓	✓	✓	✓
House Characteristics	✓	✓	✓	✓				
Block Group + Time FE	✓	✓	✓	✓				
House FE					✓	✓	✓	✓
Tract x Time FE					✓	✓	✓	✓

*p<0.1; **p<0.05; ***p<0.01

Notes: The dependent variable in every column is the log of sales price. Panel A includes all transactions where we know the race of the buyer, regardless of the number of times the house transacted. Panel B only includes houses that transacted at least twice during the study period (i.e. repeat-sales). Every column includes controls for distressed (shortsale and REO) transactions. Panel A uses a hedonic estimation that includes house characteristics, whereas Panel B uses house fixed effects. The house characteristics include continuous measures (log of age and log of square feet living area) and indicator variables (acres, bedrooms, bathrooms, garage, carport, and pool). Columns 2 to 4 are filtered by neighborhood racial composition in both panels. Column 2 includes transactions in neighborhoods that are less than 50 percent white. Column 3 (4) includes transactions in neighborhoods that are greater than or equal to 50 (80) percent white.

Table 6: Buyer racial price differentials

	Panel A: Black-white differential				Panel B: Hispanic-white differential				
	≥0.0	<0.5	≥0.5	≥0.8	≥0.0	<0.5	≥0.5	≥0.8	
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)	Obs.
Baseline	0.036***	0.059***	0.022***	0.021	0.012**	0.029	0.004	0.006	97,001
	(0.006)	(0.016)	(0.006)	(0.015)	(0.006)	(0.021)	(0.006)	(0.014)	
Owner-occupied	0.022***	0.044*	0.012**	0.011	0.004	0.021	−0.001	0.004	80,528
	(0.006)	(0.024)	(0.006)	(0.015)	(0.006)	(0.028)	(0.007)	(0.016)	
Hold > 1,095	0.016	0.038	0.001	−0.026	−0.001	−0.001	−0.004	0.013	53,919
	(0.011)	(0.066)	(0.010)	(0.024)	(0.011)	(0.081)	(0.011)	(0.028)	
No distress	0.005	0.019	−0.011	−0.038	−0.017	−0.061	−0.018	−0.011	40,811
	(0.013)	(0.078)	(0.013)	(0.030)	(0.011)	(0.073)	(0.012)	(0.025)	

*p<0.1; **p<0.05; ***p<0.01

Notes: The estimates are derived from a regression of log transaction price on a set of transaction controls, fixed effects, and race indicators. The transaction controls include indicator variables for houses that were remodeled or involved in a distressed transaction (e.g. shortsale or REO). Both house and tract by time fixed effects are included in every specification. Column 1 of Panels A and B include all repeat-sales transactions. Columns 2 to 4 are filtered by neighborhood racial composition. Column 2 includes repeat-sales transactions in neighborhoods that are less than 50 percent white. Column 3 (4) includes repeat-sales transactions in neighborhoods that are greater than or equal to 50 (80) percent white. Additional cumulative filters are applied in descending order by row as follows: Baseline includes the entire repeat-sales sample; Owner-occupied filters out investor purchases; Hold > 1,095 filters out properties that were flipped within three years or remodeled; No distress filters out all repeat-sales pairs in which at least one transaction was a distressed sale. Standard errors clustered at the house and time level are reported in brackets.

Table 7: Repeat-sales with tokens

Panel A: Unigram tokens				
	(1)	(2)	(3)	(4)
Black buyer	0.019**	0.009	0.016**	0.011
	(0.008)	(0.007)	(0.007)	(0.007)
N	35,224	35,224	35,224	35,224
R^2	0.943	0.954	0.946	0.957
Panel B: Bigram tokens				
	(1)	(2)	(3)	(4)
Black buyer	0.019**	0.009	0.015**	0.010
	(0.008)	(0.007)	(0.008)	(0.007)
N	35,224	35,224	35,224	35,224
R^2	0.943	0.949	0.945	0.950
Panel C: Flex-gram tokens				
	(1)	(2)	(3)	(4)
Black buyer	0.019**	0.010	0.017**	0.011
	(0.008)	(0.007)	(0.008)	(0.007)
N	35,224	35,224	35,224	35,224
R^2	0.943	0.954	0.946	0.956
House FE	✓	✓	✓	✓
Time FE	✓	✓	✓	✓
Time-Varying Tokens		✓		✓
Time-Invariant Tokens			✓	✓

*p<0.1; **p<0.05; ***p<0.01

Notes: The dependent variable in every column is the log of sales price. All models include a set of house and time fixed effects. Tokens are single word (unigram), two word (bigram), or multi-word (flex-gram) phrases from the MLS remarks. Column 1 includes no tokens, column 2 includes time-varying tokens, column 3 includes time-invariant tokens, and column 4 includes both time-varying and time-invariant tokens. All standard errors are two-way clustered at the house and time levels.

Table 8: Racial price differentials with financial controls

	Baseline (1)	Preapproval (2)	Income (3)	Downpayment (4)
Black buyer	0.002	0.002	0.007	0.005
	(0.014)	(0.014)	(0.014)	(0.014)
Preapproval		−0.010	−0.009	−0.010
		(0.014)	(0.014)	(0.014)
Downpayment			0.062	0.064
			(0.061)	(0.061)
Downpayment^2			0.023	0.038
			(0.101)	(0.102)
Income				0.001***
				(0.000)
Income^2				−0.000***
				(0.000)
N	39,186	39,186	39,186	39,186
R^2	0.962	0.962	0.962	0.963
House FE	✓	✓	✓	✓
Tract x Time FE	✓	✓	✓	✓

*p<0.1; **p<0.05; ***p<0.01

Notes: The estimates are derived from a regression of log transaction price on a set of fixed effects and race indicators. The repeat-sales sample used in this table does not include investor purchases, holding periods less than 3 years, distressed transactions, or transactions in which the buyer's income is unavailable. Both house and tract by time fixed effects are included in every specification. Column 1 presents a baseline estimate using the filtered repeat-sales sample. Columns 2 to 4 additively introduce a series of financial controls: preapproval, buyer income, and downpayment. Standard errors clustered at the house and time level are reported in brackets.

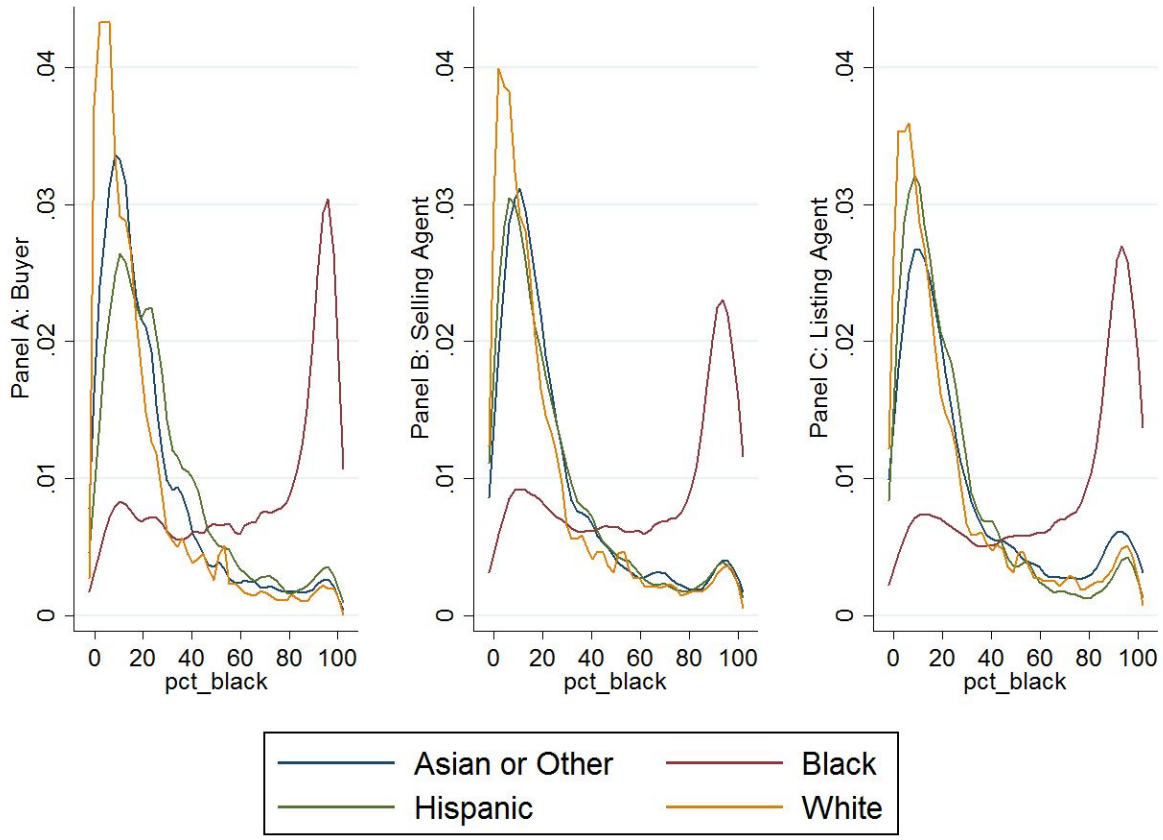
Table 9: Buyer's agent racial price differentials

	Black-white differential				
	≥0.0	<0.5	≥0.5	≥0.8	
	(1)	(2)	(3)	(4)	Obs.
Baseline	0.020	0.083	−0.033	0.075*	12,331
	(0.020)	(0.063)	(0.021)	(0.045)	
Owner-occupied	0.020	0.086	−0.026	0.076	11,788
	(0.021)	(0.068)	(0.022)	(0.047)	
Hold > 1,095	−0.039	−0.044	−0.034	0.013	8,193
	(0.025)	(0.113)	(0.025)	(0.061)	
No distress	−0.014	−0.024	−0.008	0.001	6,849
	(0.026)	(0.151)	(0.026)	(0.045)	

*p<0.1; **p<0.05; ***p<0.01

Notes: The estimates are derived from a regression of log transaction price on a set of transaction controls, fixed effects, and race indicators. The transaction controls include indicator variables for houses that were remodeled or involved in a distressed transaction (e.g. shortsale or REO). Both house and tract by time fixed effects are included in every specification. Column 1 includes all repeat-sales transactions. Columns 2 to 4 are filtered by neighborhood racial composition. Column 2 includes repeat-sales transactions in neighborhoods that are less than 50% white. Column 3 (4) includes repeat-sales transactions in neighborhoods that are greater than or equal to 50% (80%) white. Additional cumulative filters are applied in descending order by row as follows: Baseline includes the entire repeat-sales sample; Owner-occupied filters out investor purchases; Hold > 1,095 filters out properties that were flipped within three years or remodeled; No distress filters out all repeat-sales pairs in which at least one transaction was a distressed sale. Standard errors clustered at the house and time level are reported in brackets.

Figure 1: Kernel Density by Race

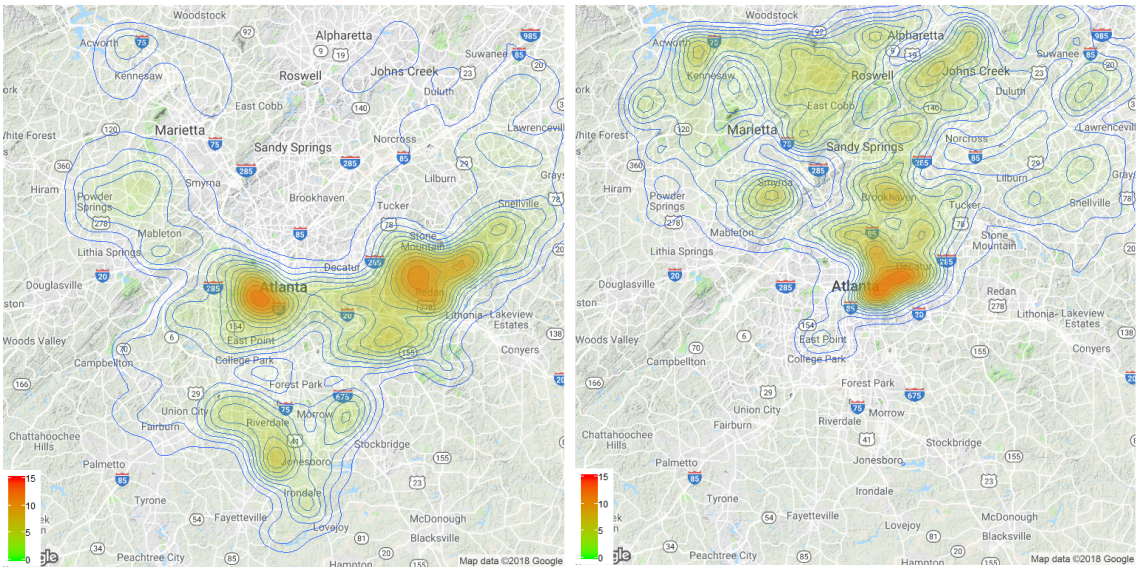


Notes: This figure plots density estimates by race and neighborhood composition for three distinct market participants. Panel A plots the transaction density by buyer race relative to the fraction of the neighborhood's population that is black. Panel B and C plot similar densities for buyer agent and listing agent race, respectively. The neighborhood's racial composition is measured at the 2010 census block group level.

Figure 2: Transactions by race in Atlanta, Georgia

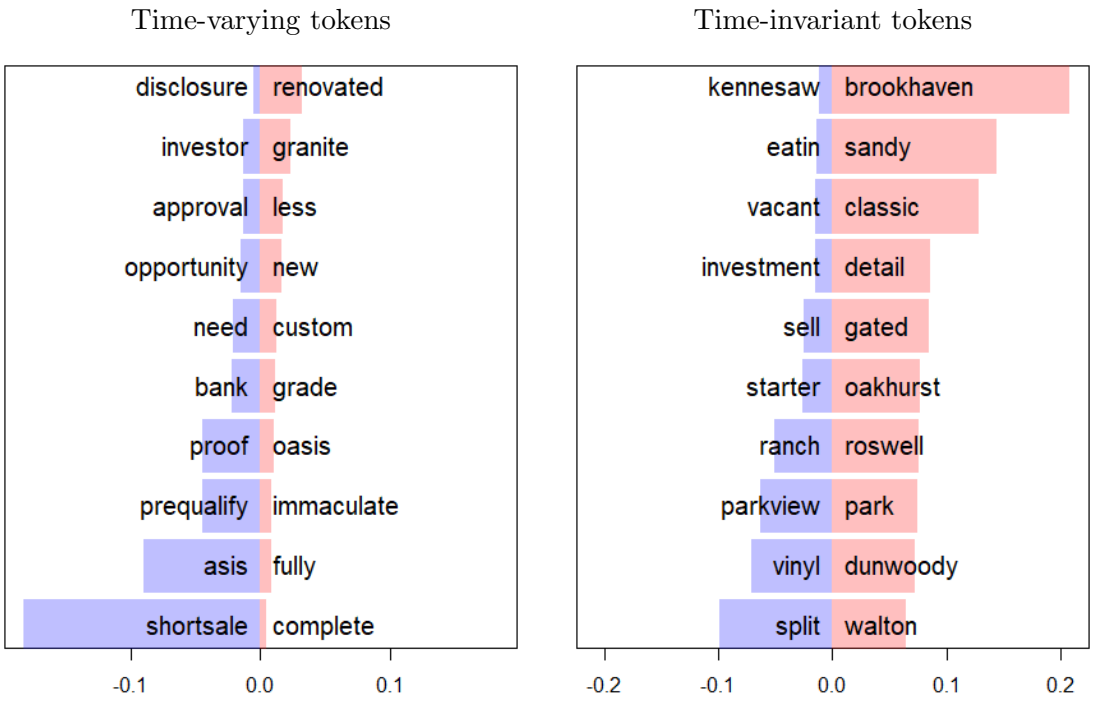
Panel A: Black buyers

Panel B: White buyers



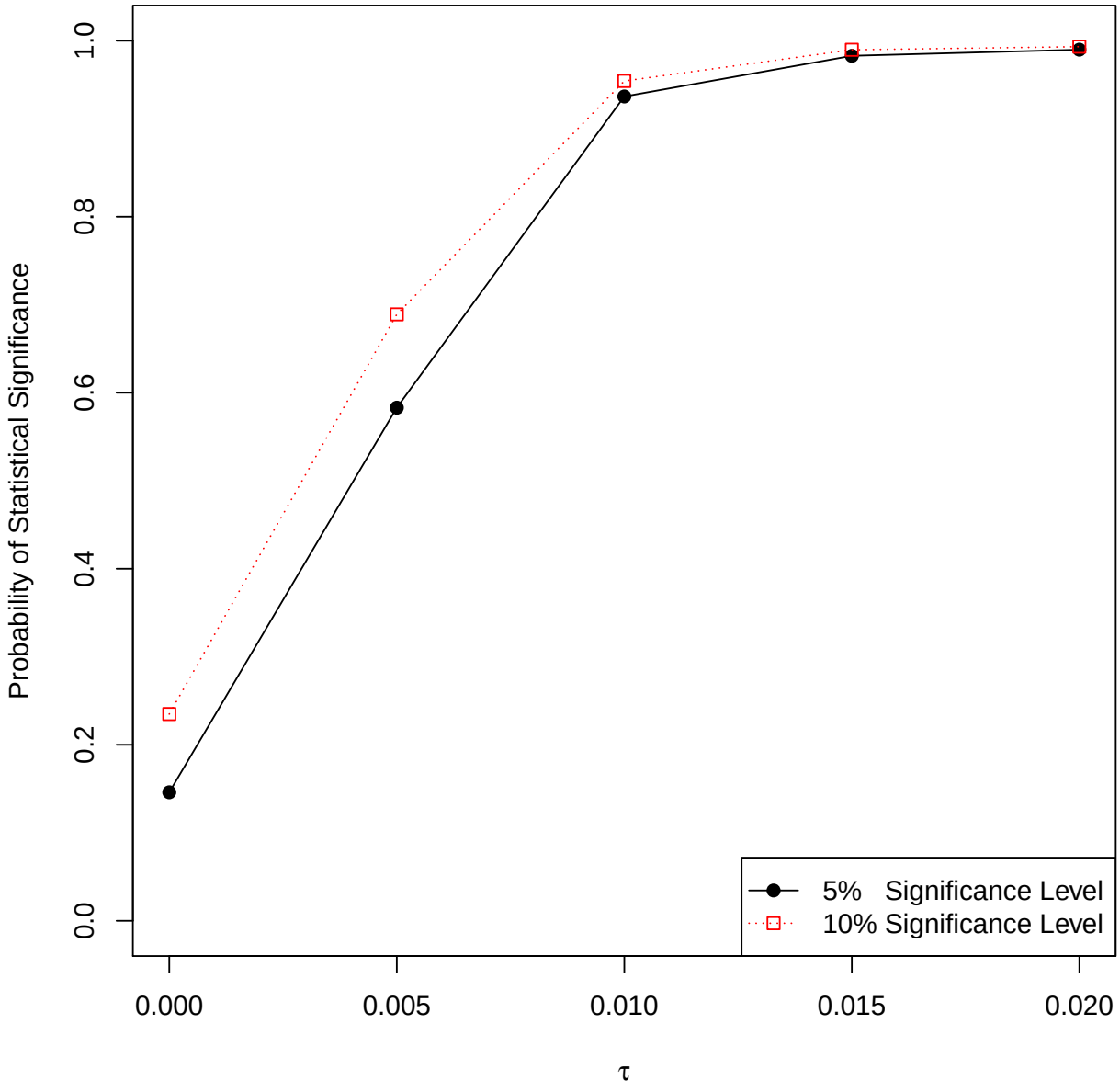
Notes: This figure plots the geographical distribution of house purchases by the race of the buyer. Panel A plots house purchases by black buyers and Panel B plots house purchases by white buyers.

Figure 3: Implicit prices for unigram tokens



Notes: The top ten positive and negative time-varying and time-invariant unigram tokens are displayed. The implicit prices displayed in the figure represent the token's coefficient from the first stage of the double-selection LASSO procedure.

Figure 4: Post Double-Selection in Finite Samples



Notes: This figure displays simulation results for the finite sample performance of the post double-selection estimator. The figure displays the fraction of simulations where $H_0 : \tau = 0$ is rejected at either a 5% or 10% significance level. Simulations use dimensions ($N = 35,224$, time fixed effects, house fixed effects, and $K = 2,000$ candidate tokens) and parameter estimates taken from the repeat-sales data set. Details of the simulation are provided in the appendix.

Appendices

Contents

A Data Overview

A.1 Data filters

To eliminate outliers, we drop transactions that do not meet the following criteria:

1. \$30,000 $\leq$ sales price $\leq$ \$3,000,000
2. 500 $\leq$ square feet of living area $\leq$ 6,000
3. 1 $\leq$ bedrooms $\leq$ 6
4. 1 $\leq$ bathrooms $\leq$ 6
5. age $\geq$ 2
6. acres $\leq$ 5

A.2 Merge process

This section describes the data preparation and merge process that combines the information from three data sources: CoreLogic, HMDA, and GAMLS. Prior to merging the three datasets we clean the raw data using the filters listed in Section A.1. After cleaning the data we use the CoreLogic data to join both the publicly available loan application registry (LAR) data collected under HMDA and the GAMLS data. Merging the GAMLS and LAR data directly is not possible.

The merge process joins the CoreLogic and LAR data using the following fields: census tract number, lender, loan amount, and year. A house's census tract assignment can change over time, so we identify the appropriate census tract number for each transaction based on the year of the transaction. Transactions prior to 2003 use their 1990 Census assignment; transactions from 2003 through 2011 use their 2000 Census assignment; and transactions from 2012 forward use their 2010 Census assignment. We use the first three characters of the lender's name to avoid dropping matches where the lender's name is abbreviated in one of the two files (e.g. *Wells Fargo Bank* versus *Wells Fargo Bank NA*).

Prior to merging the two files, we drop duplicate records based on the unique identifier that is formed using the four fields. Just under 5 percent of the transactions in the CoreLogic data are dropped because they are duplicates. Approximately 63.2 percent of the remaining sales transactions in the CoreLogic dataset are successfully matched. The match rate is comparable to Bayer et al. (2017). The resulting dataset includes 282,095 transactions. Of which, 97,001 are repeat-sales transactions.

Next we merge the GAMLS file with the CoreLogic-LAR dataset using the following fields: property address, sale date (mm/yyyy), and sales price. Duplicate records in which the house has multiple transactions within the same month are dropped. We do not require an exact match on sales price. Instead, we first match records based on the property address and sale date fields. Then we drop records in which the sales price is not within a plus or minus three percent range. The resulting dataset includes 160,844 transactions. Of which,

35,224 are repeat-sales transactions.

Descriptive statistics for the full, repeat-sales, and non-repeat-sales merged datasets are provided in Table A1. Panel A displays the descriptive statistics for every transaction in which we know the race of the buyer. Panel A represents the merged CoreLogic-LAR dataset which does not include the textual information about the house (i.e. remarks) or identify the real estate agents involved in the transaction since both datapoints are only available in the GAMLS. Panel B represents a subsample of Panel A in which we were able to successfully merge the CoreLogic-LAR data with the GAMLS data. The subsample in Panel B includes transactions in which we know the race of the buyer, regardless of whether we were able to identify the race of the real estate agents involved in the transaction. In contrast, Panel C includes transactions in which we know the race of the buyer's agent, regardless of whether we were able to identify the race of the buyer involved in the transaction.

Table A1: Descriptive statistics for the merged datasets

	Panel A: CoreLogic Buyer			Panel B: GAMLS Buyer			Panel C: Buyer's Agent		
	Full (1)	Repeat (2)	Single (3)	Full (1)	Repeat (2)	Single (3)	Full (1)	Repeat (2)	Single (3)
Observations	282,095	97,001	185,094	160,844	35,224	125,620	82,427	12,331	70,096
House characteristics									
Price (000s)	246.89	256.28	241.96	233.26	242.46	230.67	256.49	277.75	252.76
Sfla (000s)	2.32	2.29	2.34	2.30	2.29	2.30	2.38	2.36	2.39
Age	26.08	28.28	24.92	25.33	26.88	24.90	28.27	30.60	27.86
Bedrooms	3.43	3.41	3.45	3.72	3.69	3.73	3.75	3.70	3.76
Bathrooms	2.73	2.72	2.74	2.63	2.61	2.64	2.71	2.70	2.71
Owner-occupier	0.90	0.90	0.91	0.92	0.93	0.92			
Race (proportions)									
Asian and other	0.10	0.09	0.11	0.11	0.09	0.11	0.02	0.01	0.02
Black	0.26	0.22	0.29	0.26	0.19	0.28	0.12	0.08	0.12
Hispanic	0.07	0.06	0.07	0.07	0.07	0.07	0.04	0.04	0.04
White	0.57	0.63	0.54	0.56	0.65	0.53	0.83	0.88	0.82
CoreLogic	✓	✓	✓	✓	✓	✓			
LAR	✓	✓	✓	✓	✓	✓			
GAMLS				✓	✓	✓	✓	✓	✓

A.3 Paired repeat-sales transactions by race

We report racial price differentials for several filtered subsamples in Table 6. For each filtered subsample, we also report racial price differentials across differing neighborhood racial compositions in columns 1 to 4. Column 1 includes all repeat-sales transactions (percent white ≥ 0.0). Column 2 includes repeat-sales transactions in neighborhoods that are less than 50 percent white. Column 3 includes repeat-sales transactions in neighborhoods that are greater than or equal to 50 percent white and column 4 includes transactions in neighborhoods that are greater than or equal to 80 percent white.

Given that buyers sort into neighborhoods based on race, the number of black-white transactions likely differs based on the racial composition of the neighborhood. Table A2 provides a detailed breakdown of the transaction type by race for each filtered subsample. The results in Panel A show that transactions involving black and white buyers/sellers are much more likely to occur in non-majority white neighborhoods (19 percent) relative to majority white neighborhoods (10 percent). Whereas, transactions involving Hispanic and white buyers/sellers represent 4 percent of the sample in non-majority white neighborhoods and 7 percent of the sample in majority white neighborhoods.

The results in Table A2 also highlight the fact that the filters remove a higher percentage of repeat-sales transaction pairs in non-majority white neighborhoods relative to majority white neighborhoods. For example, the first filter we apply in the paper identifies and removes purchases by individuals that are not owner-occupiers. The filter removes approximately 17.6 percent of the repeat-sales pairs in column 1 of Panel B. However, a much higher percentage of repeat-sales pairs in non-majority white neighborhoods are removed relative to majority white neighborhoods (28.6 percent in column 2 versus 12.2 percent in column 3). This result suggests that investors were more active in non-majority white neighborhoods during the study period.

The second filter drops houses that were remodeled and/or held for less than three years. The filters are cumulative, so the second filter removes approximately 34.9 percent of the

repeat-sales pairs. Unlike the first filter, the effect of the second filter is fairly uniform across the neighborhood racial compositions (38.9 percent in column 2 versus 33.4 percent in column 3). The third and final filter removes all repeat-sales pairs in which at least one transaction is a distressed sale. Similar to the first filter, a much higher percentage of repeat-sales pairs in non-majority white neighborhoods are removed relative to majority white neighborhoods (38.2 percent in column 2 versus 19.5 percent in column 3).

Table A2: Paired repeat-sales transactions by race

	≥0.0 (1)	<0.5 (2)	≥0.5 (3)	≥0.8 (4)
Panel A: Baseline				
Black-to-Black	0.13	0.34	0.02	0.01
Black-to-Hispanic	0.01	0.03	0.01	0.00
Black-to-White	0.06	0.09	0.04	0.03
Hispanic-to-Black	0.01	0.01	0.00	0.00
Hispanic-to-Hispanic	0.01	0.02	0.01	0.00
Hispanic-to-White	0.02	0.01	0.02	0.02
White-to-Black	0.07	0.10	0.06	0.04
White-to-Hispanic	0.04	0.03	0.05	0.04
White-to-White	0.49	0.20	0.63	0.74
Other types	0.16	0.17	0.16	0.12
Transaction pairs	51,943	17,073	34,870	14,864
Panel B: Owner-occupied				
Black-to-Black	0.10	0.31	0.02	0.01
Black-to-Hispanic	0.01	0.03	0.01	0.00
Black-to-White	0.05	0.08	0.04	0.03
Hispanic-to-Black	0.01	0.01	0.00	0.00
Hispanic-to-Hispanic	0.01	0.02	0.01	0.00
Hispanic-to-White	0.02	0.01	0.02	0.02
White-to-Black	0.07	0.08	0.06	0.04
White-to-Hispanic	0.04	0.04	0.05	0.04
White-to-White	0.53	0.23	0.64	0.75
Other types	0.16	0.18	0.15	0.12
Transaction pairs	42,816	12,189	30,627	13,379
Panel C: Hold > 1,095				
Black-to-Black	0.09	0.28	0.02	0.00
Black-to-Hispanic	0.01	0.03	0.01	0.00
Black-to-White	0.05	0.07	0.04	0.03
Hispanic-to-Black	0.01	0.01	0.00	0.00
Hispanic-to-Hispanic	0.01	0.02	0.01	0.00
Hispanic-to-White	0.02	0.01	0.02	0.02
White-to-Black	0.07	0.09	0.06	0.04
White-to-Hispanic	0.05	0.04	0.05	0.04
White-to-White	0.54	0.25	0.65	0.75
Other types	0.16	0.19	0.15	0.12
Transaction pairs	27,860	7,451	20,409	8,907
Panel D: No Distress				
Black-to-Black	0.06	0.22	0.01	0.00
Black-to-Hispanic	0.01	0.03	0.01	0.00
Black-to-White	0.04	0.07	0.03	0.02
Hispanic-to-Black	0.00	0.01	0.00	0.00
Hispanic-to-Hispanic	0.00	0.01	0.00	0.00
Hispanic-to-White	0.01	0.01	0.02	0.02
White-to-Black	0.06	0.09	0.06	0.04
White-to-Hispanic	0.05	0.05	0.05	0.04
White-to-White	0.60	0.30	0.68	0.76
Other types	0.16	0.21	0.15	0.12
Transaction pairs	21,038	4,604	16,434	7,679

B Tokenization process

B.1 Cleaning and standardization

The tokens in this study refer to either single words (*unigrams*), two-word phrases (*bigrams*), or words or phrases that are constituent parts of larger phrases (*flex-grams*).[24] Before creating the tokens, we perform a series of steps to clean and standardize the remarks. The steps are performed in the following order:

1. Convert to lower case.
2. Replace commas (,) periods (.), ampersands (&) and the word *and* with ***" STOP "***. A space is placed at the beginning and end of ***STOP***.
3. Replace all special characters with a space.
4. Replace apostrophes.
5. Remove all remaining single letters.
6. Replace all numbers with a space. Numbers can be in either numeric or character form.
7. Remove repeated ***STOP***s and trim white space at the beginning and end.
8. Use open-source spell checking software (**hunspell**) to correct spelling mistakes.
9. Depluralize.
10. Remove a list of stop words (e.g. *a, the,* and *but*).

After cleaning the remarks, we create indicator variables based on the presence of a given token in the remarks. The $K = 2,000$ most frequent tokens form a candidate token set. However, only the $Q \ll K$ relevant tokens identified by the variable selection procedure are included in the repeat-sales specification. We futher differentiate the relevant tokens into time-varying and time-invariant token sets as follows. First, we identify the relevant tokens when house fixed effects are included. Then we identify the relevant tokens when house fixed effects are not included. By construction, the complementary tokens capture the time-invariant attributes of the house.

In unreported results, we test the sensitivity of our findings using alternative tokenization processes including, but not limited to, the use of raw remarks (i.e. uncleaned tokens) and

[24] See Liu et al. (2018) for a detailed description of the flex-gram token creation process.

stemmed tokens. Whether or not the text is preprocessed affects the number of tokens (Q_2, Q_2^*) selected by the variable selection procedure, but has no discernible affect on $\hat{\tau}_R$. In other words, the racial price differentials reported in Table 7 are not sensitive to the use of the raw text or stemmed tokens.

C Textual Analysis (Internet)

C.1 Unigram tokens

The bulk of the unigram tokens displayed in Figure 3 are self-explanatory. However, some tokens require additional explanation. This section provides a description of the top ten positive and negative time-varying and time-invariant unigram tokens displayed in Figure 3. It also provides an example of each token in a public remark.

C.1.1 Time-varying positive token descriptions

Token: **renovated**
Description: The renovated token identifies houses whose condition has been recently improved. *Example*: completely renovated bungalow w/ open front porch, hardwood floors, beautiful new bathrooms, huge dining room and kitchen which leads to a large flat backyard that is great for entertaining.
Token: **granite**
Description: The granite token identifies houses that have granite countertops. *Example*: total renovation in this 1920's candler park bungalow. kitchen has new granite countertops,viking appliances & custom cherry cabinets. master bath has granite, marble & double sinks. bring your buyers. this will not last long!
Token: **less**
Description: The less token identifies features of the house (i.e. roof, hvac, appliances, etc.) that were recently updated. *Example*: back on market! fabulous 4 sided brick ranch with tons of possibilities. home has great bones and looking for that new owner ready to make it their own. roof is less than 2 years old. come on and check it out to see the potential!
Token: **new**
Description: The new token identifies features of the house that are brand new. *Example*: outstanding home! incredible neighborhood! award winning schools! i'll try to cover most features; hardwood floors throughout, updated kitchen with new appliances and cabinets, 4 large bedrooms with a complete master suite with plenty of space and natura
Token: **custom**
Description: The custom token identifies features of the house that were built specifically for the house in question. *Example*: mint condition newer home in a prime "intown" location! huge open floorplan, hardwood floors, new carpet upstairs, new interior paint throughout, stunning master suite, custom walk in closet, professionally landscaped! excellent schools!

Token: **grade**

Description: The grade token identifies "high end" features of the house.
Example: gorgeous 3 br, 2.5 ba home in fantastic location! great open floor plan features foyer, 2-story family room, updated ceiling fans & light fixtures. wonderful open kitchen w/granite countertops & professional grade appliances. hardwood floors throughout.

Token: **oasis**

Description: The oasis token identifies houses with attractive landscaping and backyards.
Example: drop dead gorgeous! backyard oasis - come home and relax and unwind in this spectacular outdoor entertainment area! beautiful grounds include salt water pebble tec pool/spa with cascading waterfalls! spacious and open newer construction home with high-end features throughout.

Token: **immaculate**

Description: The immaculate token identifies houses in great condition.
Example: immaculate & pristine home! over 150k in upgrades! mahogany wood w/cherry stained cabinets & wood throughout. 2 story family room with stone fireplace & bookcases. bonus room that can serve as large office or 5th bedroom.

Token: **fully**

Description: The fully token identifies houses that have been completely renovated.
Example: fully renovated beautiful home. kitchen and bathroom floors upgraded to ceramic tile. new hardwood laminate flooring throughout all living areas and bedrooms. new energy saving hvac unit installed in 2014. all bathrooms have new high efficiency toilets on top of tons of other upgrades.

Token: **complete**

Description: The complete token identifies houses that have been fully renovated.
Example: complete renovation down to the studs! owner spared no expense. hardwoods throughout the living and bedroom areas, travertine tile in the kitchen and laundry room. new stainless steel appliances, granite counter tops, and kitchen island over looking fabulous gourmet kitchen.

C.1.2 Time-varying negative token descriptions

Token: **shortsale**
Description: The shortsale token identifies sellers whose mortgage balance is greater than the house's market value (i.e. distressed transactions). *Example*: approved shortsale. approved price is $112,000. very nice home in well established community. carpet and paint needed. investors and first time home buyers are welcome. please see private remarks. buyer must prequal and use closing attorney.
Token: **asis**
Description: The asis token identifies properties that are sold "as is". In other words, the seller will not make improvements to the property, so the buyer must factor the cost of the improvements into their offer. *Example*: charming all brick ranch, 2 bed rooms, formal dining room, side porch and more. in need of some tlc, sold as-is with lots of potential. must see!!
Token: **prequalify**
Description: The prequalify token typically identifies distressed transactions in which the buyer must be prequalified to make an offer. *Example*: great investment property with beautiful wood floors and kitchen. Bright dining area, 2 bedrooms, 1 full bath, a great room and a den. Foreclosure. Must prequalify with lender.
Token: **proof**
Description: The proof token typically identifies distressed transactions in which cash buyers must provide "proof of funds" prior to making an "all cash" offer. *Example*: sold as-is, where-is. no property disclosure. no termite letter. must have pre-approval letter, or proof of funds for cash, with offer. seller chooses closing attorney.
Token: **bank**
Description: The bank token identifies bank owned (i.e. distressed) transactions. These transactions are commonly referred to as real estate owned (REO) transactions. *Example*: bank owned, needs minor tlc! 3 sided brick. sold as-is, sun room, cash/cnv only, must prequalify and provide copy of em check. call for addendums.
Token: **need**
Description: The need token identifies houses in disrepair. *Example*: multi-level in need of repair/renovation. sold as-is w/ no disclosure or inspection right. prequalify or proof of funds required with offer or auto-rejected. $2000 earnest money deposit.
Token: **opportunity**
Description: The opportunity token typically identifies investment properties that need some work. *Example*: for professional renovations only, cash cow opportunity for savvy investors. only buying at $130k after repair value up to $350k great for small side project rent out in 7 days.

Token: **approval**
Description: The approval token typically identifies distressed transactions in which the buyer must already be approved for the loan to make an offer. *Example*: motivated seller submit any & all offers. priced to sale. needs minor repairs. all offers must include approval & earnest money.
Token: **investor**
Description: The investor token identifies rental properties and/or fixer uppers that can be flipped for a profit. *Example*: perfect for the investor looking for a turnkey & profitable rental property. open floor plan with ample closet and storage space. beautiful natural light fills the home. back deck and patio are perfect for weekend barbecues in expansive backyards.
Token: **disclosure**
Description: The disclosure token identifies sellers that do not provide a disclosure. This is common in distressed sales transactions. *Example*: super corporate value! inspect, compare & price for area! 3/2 frame ranch on slab, living room, bonus rm, deck - no termite or seller's disclosure.

C.1.3 Time-invariant positive token descriptions

Token: **brookhaven**
Description: The brookhaven token identifies a neighbhorhood in Atlanta. *Example*: cute home in a sought-after brookhaven location, convenient to 85, amenities, parks, and schools. hardwood floors, spacious family room, eat-in kitchen, three bedrooms, two full baths, and a sitting room pass-through off one of the bedrooms that could make a nice office.
Token: **sandy**
Description: The sandy token identifies the Sandy Springs neighborhood in Atlanta. *Example*: private mountain retreat in the heart of sandy springs! very spacious gourmet kitchen with all the bells and whistles a cook can only dream of. granite island and countertop, stainless steel appliances, large breakfast area overlooking the oversized porch.
Token: **classic**
Description: The classic token identifies houses with vintage features. *Example*: mostly renovated classic craftsman bungalow with lots of period details and character. all rooms are large with original fireplaces - 5 total. big front porch & large deck overlooking fenced in backyard.
Token: **detail**
Description: The detail token typically identifies houses that have high quality features in good condition. *Example*: charming craftsman home on quiet cul de sac! living space boasts hardwood floors & great moulding detail throughout. flowing floorplan, w/ cozy living room, separate dining, powder room, large office, & bedroom w/ ensuite bath, all on the main level. kitchen w/ granite countertops.
Token: **gated**
Description: The gated token typically identifies houses that are located in a gated community or have a gated pool. *Example*: gorgeous stepless ranch in prestigious gated community. beautiful lake front views! gourmet kitchen features granite and hardwood floors. vaulted family room with fireplace and built ins. keeping room with fireplace. new paint and new flooring, ready to move in!
Token: **oakhurst**
Description: The oakhurst token identifies a neighborhood in Atlanta. *Example*: pristine bungalow in the heart of oakhurst. just steps to the oakhurst village. large rooms, new kitchen with granite counters and stainless steel appliances. hardwood floors throughout, laundry room off of kitchen, separate room off of the master.
Token: **roswell**
Description: The roswell token identifies a neighborhood in Atlanta. *Example*: sought-after roswell subdivision convenient-four sides brick traditional situated on over an acre-updated kitchen w/ granite counters & stainless appliances - gunite pool.

Token: **park**
Description: The park token identifies houses that are located near public parks or located in a neighborhood that has the word "park" in its name. *Example*: live on one of the best streets in all of atlanta. enjoy being able to walk to shops, restaurants, piedmont park, schools, & the beltline. coveted springdale park elementary school district. welcoming large front porch overlooks picket fenced front yard.
Token: **dunwoody**
Description: The dunwoody token identifies a neighborhood in Atlanta. *Example*: absolutely beautiful home in prestigous dunwoody area. over 200k in renovations. chef's kitchen w/granite, stainless steel appliances, maple cabinets, wine fridge & sunroom/breakfast room. beautiful landscaped lot with private backyard.
Token: **walton**
Description: The walton token identifies a school district in Atlanta. *Example*: classic beauty in walton high school district! walking distance to dodgen too! living room with french doors leads to paneled den with fireplace. upgraded kitchen with granite & stainless appliances. spacious breakfast room. front & rear stairs. large bonus room.

C.1.4 Time-invariant negative token descriptions

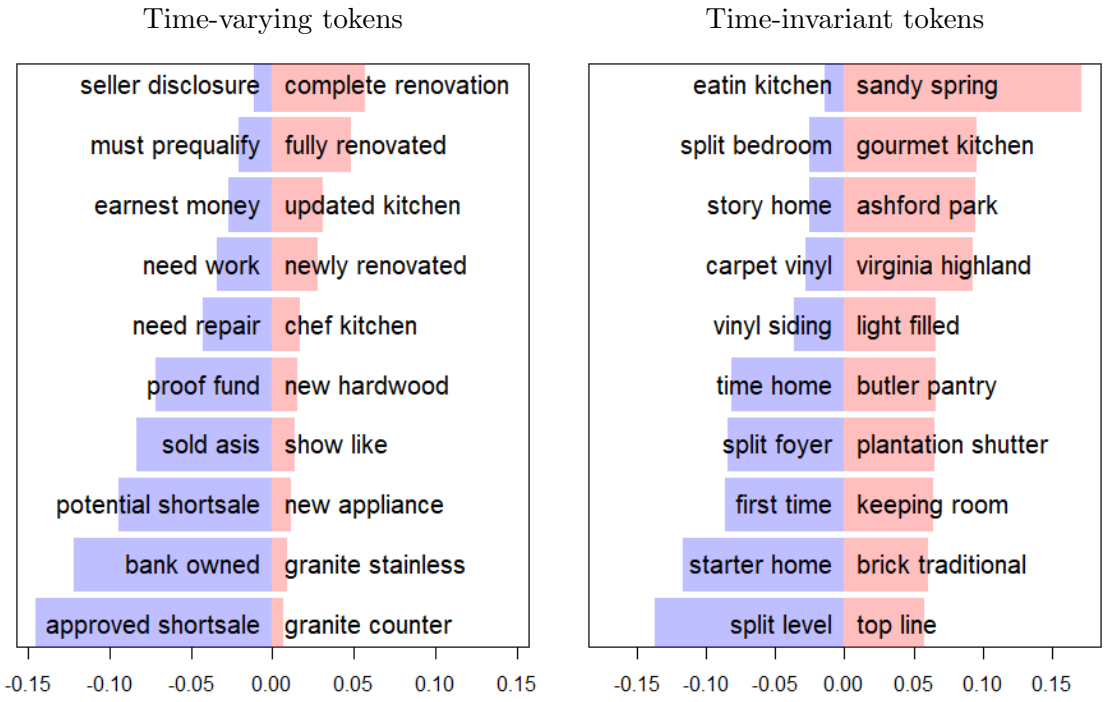
Token: **kennesaw**
Description: The kennesaw token identifies a neighborhood in Atlanta. *Example*: a beautifully maintained ranch home in a convenient kennesaw location, open floor plan, vaulted ceilings, new ceramic tile in kitchen & baths, all new lighting throughout the house.
Token: **eatin**
Description: The eatin token identifies houses with kitchens that double as dining rooms. *Example*: price reduced! motivated owners, bring all offers; clean and in good condition; master on main, whirlpool tub, walkin closets, gourmet eatin kitchen, large livingroom w/ fireplace, office w/ builtin shelfs.
Token: **vacant**
Description: The vacant token identifies houses that are unoccupied when they are listed for sale. *Example*: great schools! vacant but staged. totally renovated & stunning. custom kitchen with granite and stainless appliance. vaulted ceilings. formal living room & dining room private fully fenced.
Token: **investment**
Description: The investment token typically identifies rental (i.e. investment) properties. *Example*: investors delight instant equity, seller has apprasial, owner agent, basement finish, walk out, home warranty offered. great for investment property. currently tenant occupied with solid rental income.
Token: **sell**
Description: The sell token frequently identifies motivated sellers who must "sell their house quickly." *Example*: we must sell this house quickly!!! 10k under market value!! 4 bedroom, bonus room, incredible amount of house for the money!! 3/4 ac lot, beautifully decorated house.
Token: **starter**
Description: The starter token identifies "starter homes" that typically lack high-end features and/or are relatively small compared to the surrounding neighborhood. *Example*: this 3 bedroom 2 full bath ranch is priced for a quick sell. great starter home for small family. interior is freshly painted. owner is motivated! bring all offers!
Token: **ranch**
Description: The ranch token identifies a construction style. *Example*: seller says sell! reduced for smart buyers to get a good deal! bring all the toys - no hoa. great ranch on full basement & nearly 2 acres! brand new dishwasher & carpeting. feel out in the country, but close to everything! fabulous home is ready for first time home buyers or investor.

Token: **parkview**
Description: The parkview token identifies a neighborhood in Atlanta. *Example*: parkview ranch, split bedroom plan, vaulted open great room, separate wet bar, oversized master bedroom, walk-in closet, new carpet, new kitchen floor, wonderful screened porch.
Token: **vinyl**
Description: The vinyl token identifies houses with vinyl floors, counters and/or siding. *Example*: 4 side brick. hardwood floors throughout. plus new vinyl floors in kitchen. fenced-in backyard. view to family room & separate living room. covered back patio. large bedrooms.
Token: **split**
Description: The split token identifies houses with split foyers, levels and/or bedroom layouts. *Example*: awesome split level w/ lower level teen suite! 4bdrm/3full baths, great room w/ fireplace, eat-in kitchen, 2 car garage, level lot w/fenced backyard! mill creek schools!

C.2 Bigram tokens

This section presents the top positive and negative time-varying and time-invariant bigram tokens for the repeat-sales sample employed in Panel B of Table 7. The implicit prices displayed in Figure C1 represent the token's coefficient from the first stage of the double-selection LASSO procedure which, by design, shrinks the token's coefficient towards zero. A description and example of the bigram tokens are available from the authors by request.

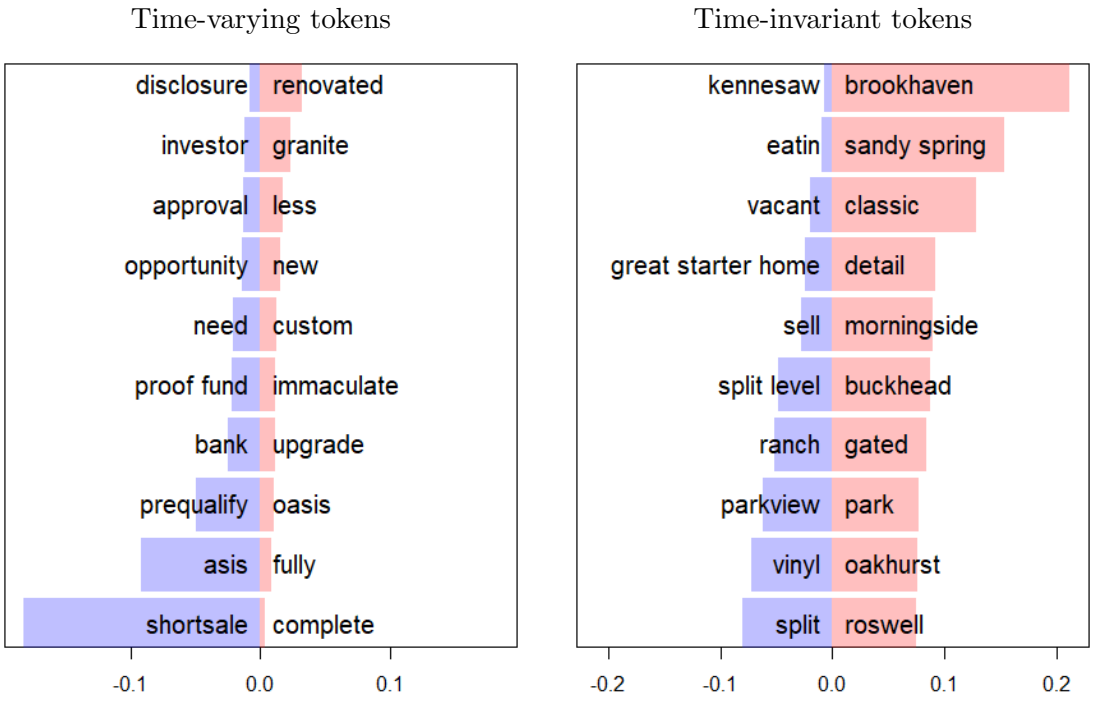
Figure C1: Implicit prices for bigram tokens



C.3 Flex-gram tokens

This section presents the top positive and negative time-varying and time-invariant flex-gram tokens for the repeat-sales sample employed in Panel C of Table 7. The implicit prices displayed in Figure C2 represent the token's coefficient from the first stage of the double-selection LASSO procedure which, by design, shrinks the token's coefficient towards zero. A description and example of the flex-gram tokens are available from the authors by request.

Figure C2: Implicit prices for bigram tokens



D Agency Relationships (Internet)

D.1 Agency Relationships

Under the laws of agency as codified for real estate agents in many states today, the listing agent exclusively represents the best interests of one client, the seller. The relationship between the buyer and their agent is less clear. An agent may represent the buyer through one of the following arrangements: buyer agency, designated agency, dual agency, or subagency.

Buyer agency occurs when the buyer signs an agreement, commonly known as a buyer brokerage agreement, which stipulates that the real estate agent will represent the buyer's best interests in locating, negotiating, and purchasing a house. Although the agent represents the buyer, they are almost always compensated by the seller. This brokerage relationship is commonly referred to as buyer agency.

In some transactions the agent representing the buyer and seller work for the same broker or brokerage firm. In which case, the broker may allow each agent to exclusively represent their respective clients. This brokerage relationship is commonly referred to as designated agency and the transactions are referred to as in-house transactions.

In Georgia, the law allows one agent to represent both the seller and the buyer as long as the agent gets the written consent of both parties. In which case, neither party is exclusively represented by the agent. This brokerage relationship is commonly referred to as dual agency. The final form of agent representation occurs when an agent works with the buyer, but represents the seller. This brokerage relationship is commonly referred to as subagency. Subagency relationships between real estate agents in Georgia, which were once the norm, are very uncommon today.

Discussions with real estate agents and brokers in the Atlanta area confirmed that subagency is not a concern during our study period. Thus, we assume the relationship between the buyer and their agent is governed by buyer agency. For this reason, we refer to the agent representing the buyer as the buyer's agent. This is necessary when using the CoreLogic-

LAR data since agent information is not available. However, the GAMLS data includes fields that identify in-house and dual agency transactions. The inclusion of these fields when using the GAMLS data does not have a discernible effect on the results we report.

D.2 For sale by owner (FSBO)

Homeowners can sell their house without the services of a listing agent. These sales are known as *For Sale by Owner* (FSBO). According to NAR (2017), the primary reasons FSBO sellers chose to sell their house without an agent were because they did not want to pay a sales commission (43%), sold to a friend or relative (23%), or were approached directly by a buyer (10%). When taken at face value, these statistics suggest that FSBO sellers are profit maximizing and do not choose FSBO to discriminate against buyers.

Comparing prices for FSBO and non-FSBO transactions could identify price effects attributable to real estate agents, ceteris parebis. However, FSBO transactions are, by definition, not included in MLS data and there is no straightforward way to identify them in the CoreLogic data. Even if we could identify FSBO transactions there are at least two concerns we would have to account for. First, a large fraction of FSBO transactions are not arms length according to the NAR (2017) survey. If the non-arms length transactions are not flagged in the FSBO data, then it is not clear how to interpret the composite estimate. Second, we would not know whether the buyer was represented by a buyer's agent. Just because the seller chose to forgo agent representation, does not mean that the buyer made the same choice. In which case, the buyer and seller may not interact or meet until the closing, so the buyer's race would not be revealed and the seller could not discriminate against them.

E Robustness Checks (Internet)

E.1 Flipper covariate in place of filters

Table E1 reports black-white and Hispanic-white price differentials using the repeat-sales approach. Instead of filtering out houses that were (i) flagged in the tax assessor data as having undergone a major renovation or (ii) held for less three years or less, we include a "flipped" indicator variable in the repeat-sales estimation. The results suggest that black buyers pay approximately 1.6 percent more than white buyers for comparable housing; this estimate is marginally statistically significant at the 10% level. However, the racial price differentials are statistically insignificant in columns 2 to 4 at conventional significance levels. We suspect that the flipped variable does not perfectly control for heterogeneous amounts of renovation that likely takes place across the housing sample. For this reason, we argue that the textual analysis approach we employ is more appropriate.

Table E1: Filtered repeat-sales with flipper covariate

	≥0.0	<0.5	≥0.5	≥0.8
	(1)	(2)	(3)	(4)
Black buyer	0.016*	0.044	0.005	−0.005
	(0.008)	(0.052)	(0.009)	(0.020)
Hispanic buyer	−0.007	−0.003	−0.012	0.001
	(0.009)	(0.063)	(0.009)	(0.019)
Flipped	0.028***	0.017	0.027***	0.031***
	(0.005)	(0.054)	(0.005)	(0.010)
Flipped x Black buyer	0.013	0.059	0.003	0.037
	(0.020)	(0.092)	(0.021)	(0.051)
Flipped x Hispanic buyer	0.012	0.059	0.004	−0.025
	(0.021)	(0.128)	(0.024)	(0.057)
N	60,263	14,150	46,113	21,383
R^2	0.992	0.994	0.992	0.992

*p<0.1; **p<0.05; ***p<0.01

Notes: The estimates are derived from a regression of log transaction price on a set of fixed effects, race indicators, and a variable that identifies if the house was flipped. Houses that were involved in a distressed transaction are not included. The flipped variable identifies house that were remodeled or held for three years or less. Both house and tract by time fixed effects are included in every specification. Columns 2 to 4 are filtered by neighborhood racial composition (percent white). Column 2 includes repeat-sales transactions in neighborhoods that are less than 50 percent white. Column 3 (4) includes repeat-sales transactions in neighborhoods that are greater than or equal to 50 (80) percent white. Standard errors clustered at the house and time level are reported in brackets.

E.2 Pre- versus post-crisis

Table E2 reports racial price differentials using the repeat-sales approach for the filtered subsamples in Table 6. In this table we interact the buyer race indicators with a post-crisis indicator variable that identifies transactions that sold after 2007. The baseline estimates suggest that a large racial price differential existed prior to the crisis, but it became economically insignificant during the post-crisis period (6.0 percent pre-crisis versus -0.4 percent post-crisis for black buyers). However, as the data is filtered the magnitude of the racial price differential decreases until it is statistically insignificant in column 4. The large disparity between the pre-crisis estimates in column 1 relative to column 4 provides additional evidence that the estimates are biased by time-varying improvements to the house that are more likely to occur during "good" times (i.e. pre-crisis) versus "bad" times (i.e. post-crisis).

Table E2: Pre- versus post-crisis racial price differentials

	Filtered Subsamples			
	Baseline	Owner-occupied	Hold $\geq$ 1,095	No Distress
	(1)	(2)	(3)	(4)
Black buyer	0.060***	0.051***	0.039***	0.023
	(0.007)	(0.008)	(0.015)	(0.021)
Hispanic buyer	0.030***	0.025***	0.021	−0.011
	(0.008)	(0.009)	(0.017)	(0.023)
Post x Black buyer	−0.064***	−0.066***	−0.044**	−0.034
	(0.011)	(0.012)	(0.018)	(0.027)
Post x Hispanic buyer	−0.037***	−0.040***	−0.036	−0.010
	(0.011)	(0.013)	(0.023)	(0.029)
N	97,001	80,528	53,919	40,811
R^2	0.985	0.989	0.994	0.996

*p<0.1; **p<0.05; ***p<0.01

Note: The estimates are derived from a regression of log transaction price on a set of transaction controls, fixed effects, and race indicators. The transaction controls include indicator variables for houses that were remodeled or involved in a distressed transaction (e.g. shortsale or REO). Both house and tract by time fixed effects are included in every specification. The columns differ only in the repeat-sales subsample used in their estimation. Column 1 uses the baseline repeat-sales sample, column 2 uses the owner-occupied filtered repeat-sales subsample, column 3 uses the hold > 1,095 filtered repeat-sales subsample, and column 4 uses the no distress filtered repeat-sales subsample. Standard errors clustered at the house and time level are reported in brackets.

E.3 Neighborhood income

The racial price differentials in column 1 of Table 6 are identical to the estimates in column 1 of Table E3 since both tables use the entire repeat-sales sample. In columns 2 and 3, we examine whether black-white and Hispanic-white price differentials vary across income levels. Transactions that take place in census tracts that have an average income below (above) the weighted average income in Atlanta are included in column 2 (3). After applying the cumulative filters to remove investor purchases, flipped properties, and distressed transactions the racial price differentials are statistically insignificant.

Table E3: Buyer racial price differentials using neighborhood income

	Panel A: Black-white differential			Panel B: Hispanic-white differential			
	All	Low	High	All	Low	High	
	(1)	(2)	(3)	(1)	(2)	(3)	Obs.
Baseline	0.036***	0.051***	0.029***	0.012**	0.016	0.010*	97,001
	(0.006)	(0.018)	(0.005)	(0.006)	(0.019)	(0.006)	
Owner-occupied	0.022***	0.030	0.020***	0.004	0.002	0.004	80,528
	(0.006)	(0.028)	(0.005)	(0.006)	(0.026)	(0.006)	
Hold > 1,095	0.016	0.056	0.008	−0.001	0.005	−0.001	53,919
	(0.011)	(0.065)	(0.009)	(0.011)	(0.057)	(0.010)	
No distress	0.005	0.015	0.004	−0.017	−0.045	−0.014	40,811
	(0.013)	(0.096)	(0.012)	(0.014)	(0.094)	(0.013)	

*p<0.1; **p<0.05; ***p<0.01

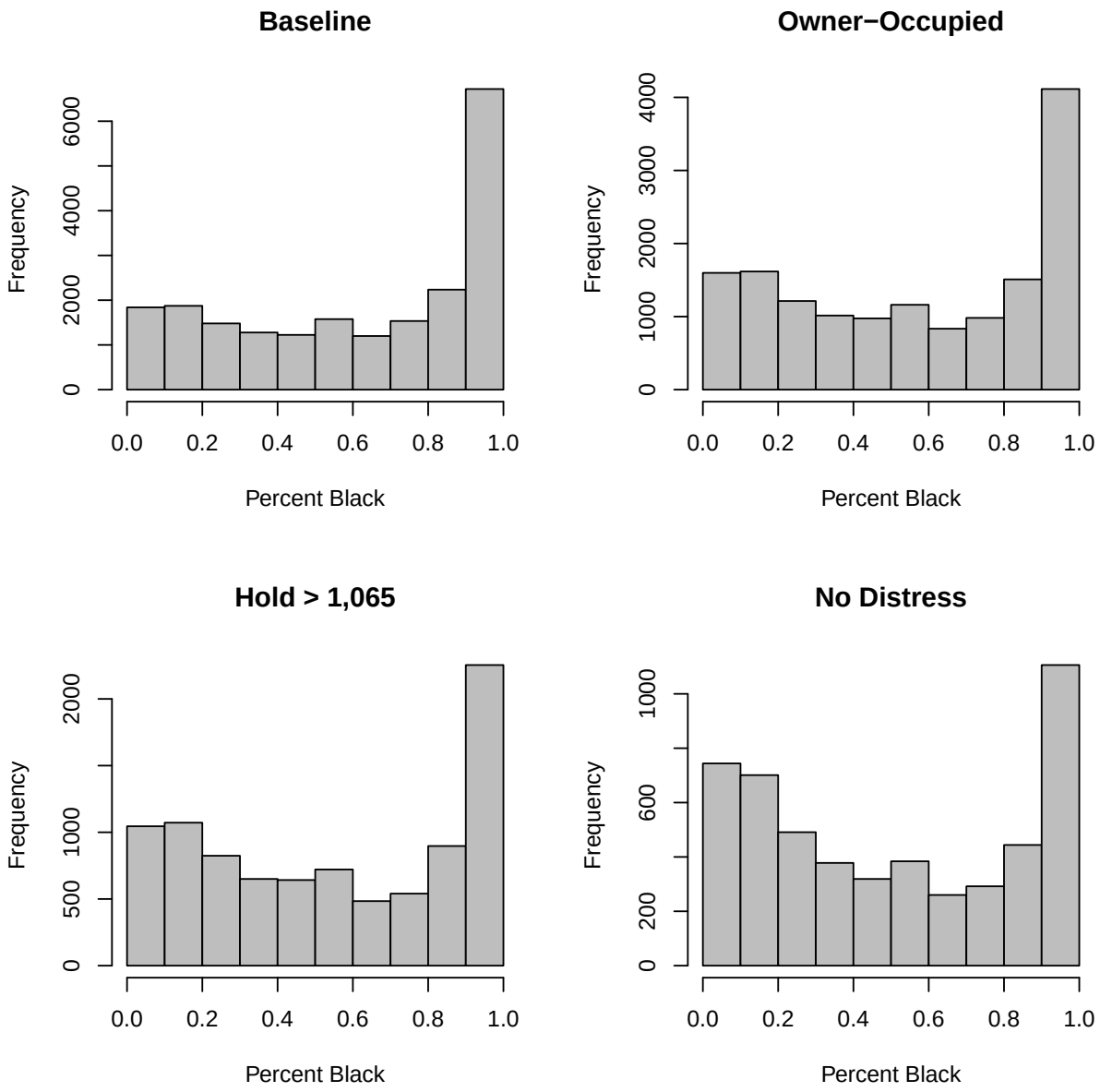
Notes: The estimates are derived from a regression of log transaction price on a set of transaction controls, fixed effects, and race indicators. The transaction controls include indicator variables for houses that were remodeled or involved in a distressed transaction (e.g. shortsale or REO). Both house and tract by time fixed effects are included in every specification. Column 1 of Panels A and B include all repeat-sales transactions. Columns 2 to 4 are filtered based on the median income of the neighborhood. Column 2 (3) includes only repeat-sales transactions in neighborhoods that have a below (above) average income. Additional cumulative filters are applied in descending order by row as follows: Baseline includes the entire repeat-sales sample; Owner-occupied filters out investor purchases; Hold > 1,095 filters out properties that were flipped within three years or remodeled; No distress filters out all repeat-sales pairs in which at least one transaction was a distressed sale. Standard errors clustered at the house and time level are reported in brackets.

E.4 Alternative neighborhood compositions

The neighborhood racial compositions in Table 6 are based on the fraction of the population that is white at the census block group level. Instead of using the fraction of the neighborhood that is white, Figure E1 presents the number of black buyers based on the fraction of the neighborhood that is black. The figure presents four histograms that differ only in their application of the cumulative sequential filters in Table 6: owner-occupied, holding period $>$ 1,095, and no distressed transactions. The bins in each histogram have a width of 0.1 and correspond to the fraction of black residents in each decile. After applying the filters the resulting subsamples have a larger fraction of black buyer transactions in neighborhoods with a smaller fraction of black residents. Alternatively, the baseline data set contains more black buyer transactions in black neighorhoods. We provide similar histograms using the fraction of white and minority (non-white) residents in Figures E2 and E3, respectively.

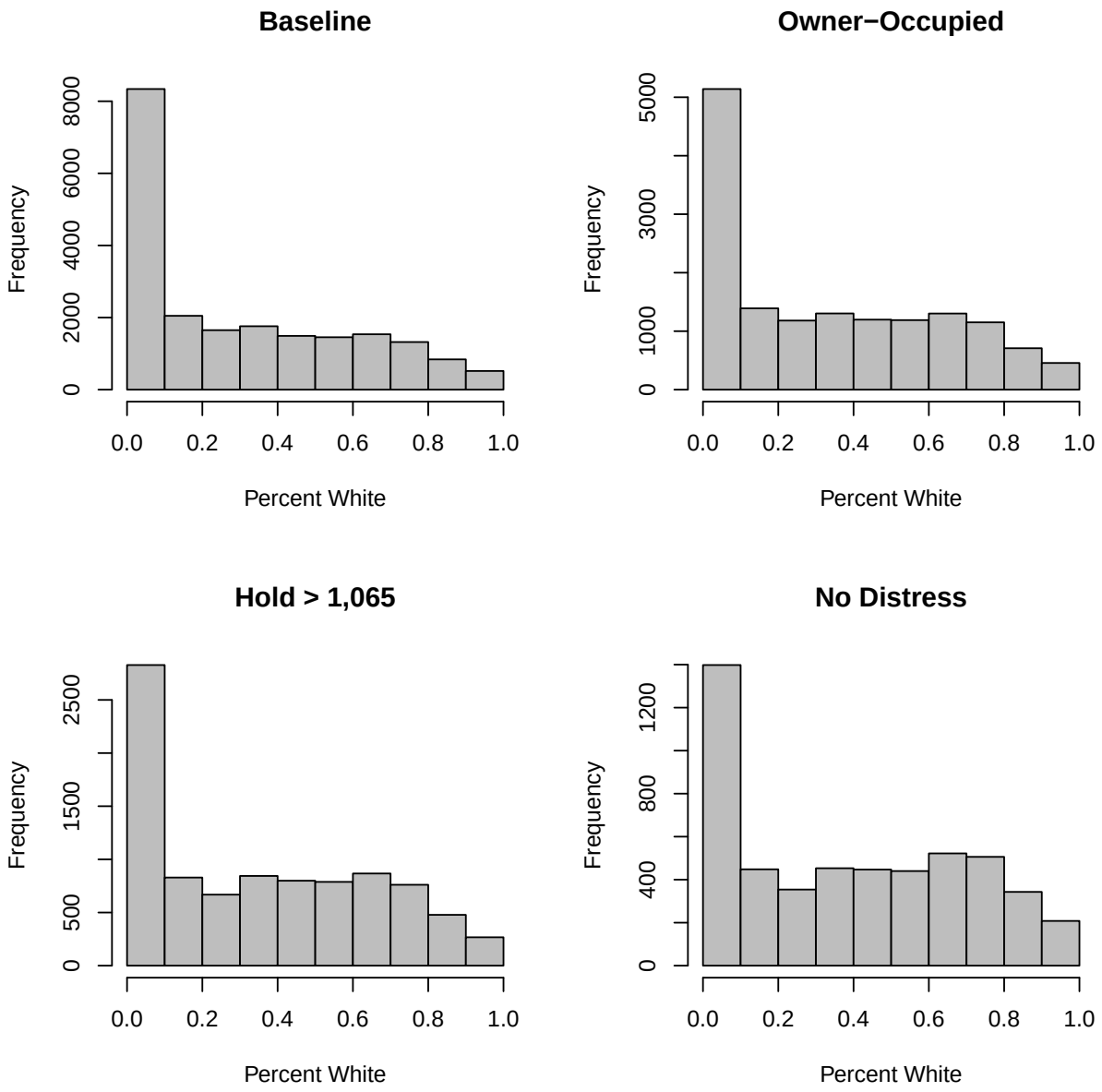
Table E4 reports black-white and Hispanic-white price differentials similar to Table 6 except that the neighborhood racial compositions in columns 2 to 4 are delineated using the fraction of the neighbhorhood that is black (instead of white). Column 1 includes all repeat-sales transactions (percent black $\geq$ 0.0). Column 2 includes repeat-sales transactions in neighborhoods that are greater than 50 percent black. Column 3 includes repeat-sales transactions in neighborhoods that are less than or equal to 50 percent black and column 4 includes transactions in neighborhoods that are less than or equal to 80 percent black. The racial price differentials in column 1 of Tables 6 and E4 are identical since they both use the entire repeat-sales sample. However, the racial price differentials in columns 2 to 4 differ slightly, suggesting that the coefficient estimates are sensitive to the racial composition filters (percent black versus percent white) used to form the subsample estimates. Critically, the racial price differentials remain statistically insignificant after the cumulative filters are applied.

Figure E1: Black buyers by fraction black



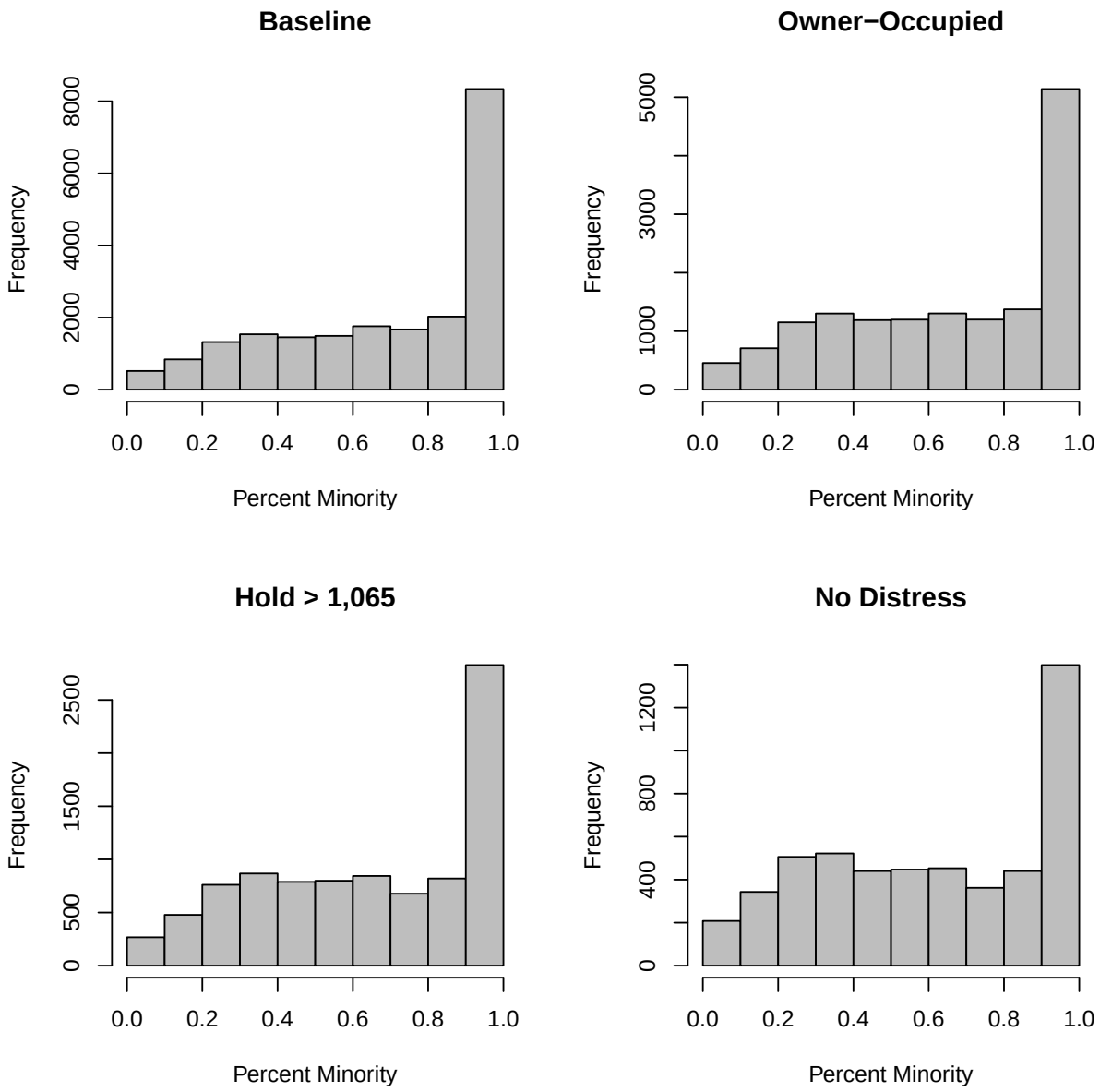
Notes: This figure displays the number of black buyers based on the fraction of black residents in the census block group. Bins correspond to the deciles $[0, 0.1], [0.1, 0.2], ..., [0.9, 1]$. Each histogram is tabulated using a series of cumulative sequential filters for owner-occupied, holding period $>$ 1,095, and no distressed transactions.

Figure E2: Black buyers by fraction white



Notes: This figure displays the number of black buyers based on the fraction of white residents in the census block group. Bins correspond to the deciles $[0, 0.1], [0.1, 0.2], ..., [0.9, 1]$. Each histogram is tabulated using a series of cumulative sequential filters for owner-occupied, holding period $>$ 1,095, and no distressed transactions.

Figure E3: Black buyers by fraction minority



Notes: This figure displays the number of black buyers based on the fraction of non-white residents in the census block group. Bins correspond to the deciles $[0, 0.1], [0.1, 0.2], ..., [0.9, 1]$. Each histogram is tabulated using a series of cumulative sequential filters for owner-occupied, holding period $>$ 1,095, and no distressed transactions.

Table E4: Buyer racial price differentials using fraction black

	Panel A: Black-white differential				Panel B: Hispanic-white differential				
	≥0.0 (1)	>0.5 (2)	≤0.5 (3)	≤0.8 (4)	≥0.0 (1)	>0.5 (2)	≤0.5 (3)	≤0.8 (4)	Obs.
Baseline	0.036***	0.065***	0.023***	0.031***	0.012**	0.014	0.009	0.011**	97,001
	(0.006)	(0.021)	(0.005)	(0.005)	(0.006)	(0.041)	(0.005)	(0.006)	
Owner-occupied	0.022***	0.049	0.015***	0.020***	0.004	0.011	0.003	0.003	80,528
	(0.006)	(0.037)	(0.005)	(0.006)	(0.006)	(0.068)	(0.006)	(0.006)	
Hold > 1,095	0.016	0.030	0.006	0.014	−0.001	−0.093	−0.000	−0.001	53,919
	(0.011)	(0.126)	(0.010)	(0.011)	(0.011)	(0.283)	(0.010)	(0.010)	
No distress	0.005	−0.014	−0.002	0.004	−0.017	−0.036	−0.016	−0.017	40,811
	(0.013)	(0.202)	(0.013)	(0.013)	(0.011)	(0.065)	(0.013)	(0.014)	

*p<0.1; **p<0.05; ***p<0.01

Notes: The estimates are derived from a regression of log transaction price on a set of transaction controls, fixed effects, and race indicators. The transaction controls include indicator variables for houses that were remodeled or involved in a distressed transaction (e.g. shortsale or REO). Both house and tract by time fixed effects are included in every specification. Column 1 of Panels A and B include all repeat-sales transactions. Columns 2 to 4 are filtered by neighborhood racial composition. Column 2 includes only repeat-sales transactions in neighborhoods that are greater than 50 percent black. Column 3 (4) includes only repeat-sales transactions in neighborhoods that are less than or equal to 50 (80) percent black. Additional cumulative filters are applied in descending order by row as follows: Baseline includes the entire repeat-sales sample; Owner-occupied filters out investor purchases; Hold > 1,095 filters out properties that were flipped within three years or remodeled; No distress filters out all repeat-sales pairs in which at least one transaction was a distressed sale. Standard errors clustered at the house and time level are reported in brackets.

E.5 Tuning parameters

In the body of the paper, we use the penalty parameters (c= 1.10 and $\gamma = 0.10$) suggested by Belloni et al. (2014) and Belloni et al. (2016) to select the time-varying and time-invariant tokens. Implicitly, c and γ determine $\hat{S}_2$, $\hat{S}_2^*$, $\hat{Q}_2$, $\hat{Q}_2^*$, $\hat{\tau}_H$,and $\hat{\tau}_R$ by controlling the amount of regularization that prevents overfitting. Here, we test the sensitivity of these estimands to the choice of c and γ by considering the following combinations of $c \in \{1.01, 1.05, 1.10\}$ and $\gamma \in \{0.01, 0.10, 0.25\}$.

We use the same repeat-sales dataset employed in Table 7 to estimate the racial price differentials in Columns 2-4 of Table 7. The results in the table below show that the penalty parameters affect the number of time-varying ($\hat{Q}_2^*$) and time-invariant ($\hat{Q}_2 - \hat{Q}_2^*$) tokens selected in columns 3 and 5, but do not have a material impact on the racial price differentials in columns 4 and 6.

Table E5: Racial price differential estimates with varying penalty parameters

Penalty		Time-varying		Time-invariant	
c	γ	$\hat{Q}_2^*$	$\hat{\tau}_R$	$\hat{Q}_2 - \hat{Q}_2^*$	$\hat{\tau}_R$
(1)	(2)	(3)	(4)	(5)	(6)
1.05	0.01	39	0.010	168	0.016**
1.10	0.01	39	0.010	157	0.015**
1.25	0.01	30	0.009	127	0.016**
1.05	0.10	45	0.009	209	0.014**
1.10	0.10	45	0.009	192	0.016**
1.25	0.10	39	0.010	154	0.016**
1.05	0.25	49	0.009	225	0.014*
1.10	0.25	46	0.009	212	0.014**
1.25	0.25	39	0.010	168	0.016**
House FE		✓	✓		
Time FE		✓	✓		
Controls				✓	✓
Tract x Time FE				✓	✓

*p<0.1; **p<0.05; ***p<0.01

F Simulations (Internet)

F.1 Simulation Details

This section describes the simulation experiments we use to calculate the power of the double-selection procedure in Belloni et al. (2014) for $\tau \in \{0.00, 0.05, 0.010, 0.015, 0.020\}$. We use $J = 500$ simulations for each choice of τ. We run the simulations for the repeat-sales data using time fixed effects, house fixed effects, and $K = 2,000$ candidate tokens.

Price parameters for simulation

In order to simulate price data, we estimate parameters of the price equation using the heteroscedastic LASSO

$$(\hat{\beta}_p', \hat{\tau}_p, \hat{\theta}_p')' = \arg\min_{\beta,\tau,\theta} \sum (p_{nt} - x_{nt}\beta - b_{nt}\tau - w_{nt}\theta)^2 + \lambda_p \sum_k |\theta_k \phi_{p,k}| \tag{10}$$

Because LASSO coefficient estimates are shrunk towards 0, we then calculate the post-LASSO (Belloni and Chernozhukov, 2013) estimates as the least-squares coefficients when regressing p_{nt} on b_{nt} and the $\hat{Q}_p$ variables in x_{nt} and w_{nt} corresponding to the non-zero elements in $\hat{\beta}_p$ and $\hat{\theta}_p$. Define the $\hat{\beta}_p^{PL}$ and $\hat{\theta}_p^{PL}$ as the post-LASSO estimates and x_{nt}^{PL} and w_{nt}^{PL} as the corresponding regressors. Define $\hat{\tau}^{PL}$ as the post-LASSO estimate of τ. Define $\hat{e}_{nt}^{PL}$ as the residual from the post-LASSO estimator. We calculate the predicted value of price excluding the minority effect as $\hat{p}_{nt} = x_{nt}^{PL}\hat{\beta}_p^{PL} + w_{nt}^{PL}\hat{\theta}_p^{PL}$.

Parameters for simulation

In order to simulate binary variables for minority buyers, we estimate an ℓ_1 penalized logit model for b_{nt}

$$(\hat{\beta}_b', \hat{\theta}_b')' = \arg\max_{\beta,\theta} \prod_{nt} \Lambda(x_{nt}\beta, w_{nt}\theta)^{b_{nt}} (1 - \Lambda(x_{nt}\beta, w_{nt}\theta))^{1-b_{nt}} - \lambda_b \sum_k |\theta_k| \tag{11}$$

In Equation 11, Λ is the logistic cdf, and the objective function is a penalized likelihood model. We use the λ_b that minimizes the 5-fold cross-validated likelihood. Simulation results

are not sensitive to other choices of λ_b near this choise of λ and are available upon request. We then calculate the (unpenalized) maximum likelihood estimates, $\hat{\beta}_b^{MLE}$ and $\hat{\theta}_b^{MLE}$, for the logit model using the variables corresponding to the non-zero elements in $\hat{\beta}_b'$ and $\hat{\theta}_b'$. Based on the maximum likelihood estimates, we then calculate $\hat{\pi}_{nt} = \Pr(b_{nt} = 1|x_{nt}, w_{nt}, \hat{\beta}_b^{MLE}, \hat{\theta}_b^{MLE})$ using the logistic function.

Simulation

For each simulation $j = 1, ..., J$ and each τ, we generate y_{nt}^j and b_{nt}^j as

1. Draw b_{nt}^j as a Bernoulli random variable with $\Pr(b_{nt}^j = 1) = \hat{\pi}_{nt}$
2. Draw $\epsilon_{nt}^j \sim \mathcal{N}(0, 1)$ and create $p_{nt}^j = \hat{p}_{nt} + b_{nt}^j \tau + \epsilon_{nt}^j \hat{e}_{nt}^{PL}$

The first step ensures b_n^j is a binary variable. The second step is similar to the wild bootstrap and allows for heteroscedasticity in the simulated errors. Note, $\hat{p}_{nt}$ is the predicted price minus the post double-selection estimate $\hat{\tau}^{PL}$. Based on the simulated data, we then estimate τ using the post double-selection procedure outlined in the paper. For each τ, we record the fraction of simulations where $H_0 : \tau = 0$ is rejected at the 5% or 10% level. Results are reported below.

Simulation results

Table F1: Power simulations for repeat-sales with tokens

τ	$\text{Reject}_{5\%}$	$\text{Reject}_{10\%}$
0.000	0.081	0.171
0.005	0.670	0.758
0.010	0.919	0.965
0.015	0.989	1.000
0.020	1.000	1.000

Notes: This table displays the fraction of simulations where the post double-selection estimator for the repeat-sales estimator with tokens identifies a significant τ for $\tau \in \{0.005, ..., 0.020\}$. For each τ, 500 simulations are performed. $\text{Reject}_{5\%}$ and $\text{Reject}_{10\%}$ display the fraction of simulations where $H_p : \tau = 0$ is rejected at the 5% or 10% significance level, respectively. The size and power of the test are not equal for $\tau = 0$ because the post double-selection estimator uses a linear probability model to approximate the non-linear logistic function $\hat{S}_b^{*s}$.